The Dissertation Committee for Jie Lu Certifies that this is the approved version of the following dissertation:

Investigating the Role of PolIV-siRNAs in *Arabidopsis* Seed Development

Committee:

Zengjian J. Chen, Supervisor

Enamul Huq

Christopher Sullivan

Sibum Sung

Claus Wilke

Investigating the Role of PolIV-siRNAs in *Arabidopsis* Seed Development

by

Jie Lu, B.Agriculture; M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2013

Dedication

Dedicated to my parents, Xiaotong Lu and Wenyan Wang.

Acknowledgements

First and foremost, I would like to thank my advisor Prof. Z. Jeffrey Chen for providing me a thorough and strict academic training in thinking, writing and speaking. I have been greatly influenced by his work ethics and technical passion. His overall enthusiasm and integral view on research and his great personality have made a deep impression on me. I owe him lots of gratitude for demonstrating this way of doing research. He could not even realize how much I have learned from him. Without his support and invaluable guidance, my Ph.D. study could not have been successfully completed.

I would like to thank the other members on my Ph.D. committee, Prof. Enamul Huq, Prof. Christopher Sullivan, Prof. Sibum Sung and Prof. Claus Wilke, for their guidance and encouragement. I would also like to express my gratitude to Prof. Sir. David Baulcombe, who provided enlightening advice as well as new perspectives on my research.

I thank all members of Chen lab for their love and help. Special thanks go to Dr. Changqing Zhang. It was such a great experience to work with him. I would also like to thank Marisa Miller, Dr. Xueying Guan, Dr. Helen Chen, Dr. Xiaoli Shi, Daekwan Ko and previous lab members Dr.Misook Ha, Dr. Eundok Kim, Dr. Danny Ng and Dr. Mingxiong Pang for their generous help and warm friendship.

Most importantly, none of this work would have been possible without my husband, Yaping Zhan, and my little one, Irene Zhan. Their continuous support, love and patience help me through the difficulties. Their love keeps my life in proper perspective and balance.

Investigating the Role of PolIV-siRNAs in *Arabidopsis* Seed Development

Jie Lu, Ph.D.

The University of Texas at Austin, 2013

Supervisor: Z. Jeffrey Chen

In plants and animals small interfering RNAs (siRNAs) mediate epigenetic inheritance of heterochromatin and genome stability. PolIV-siRNAs (p4-siRNAs) protect plant genome from invasion or reactivation of transposable elements (TEs). p4-siRNAs are most abundantly expressed in endosperm during early seed development and specifically from maternal genome. The role of p4-siRNAs in gene regulation during *Arabidopsis* seed development is still elusive. In this dissertation, I investigated the epigenetic regulation of gene expression by p4-siRNAs in *Arabidopsis* endosperm. First, I found maternal p4-siRNAs are regulators of parental genome imbalance and gene expression in *Arabidopsis* seeds. I analyzed high-throughput sequencing data of small RNAs from developing seeds with different parental genome contributions. I found levels of maternal siRNAs expression responded to the dosage of maternal genome dosage. I identified TE-associated genes (*TAG*s) that were associated with maternal p4-siRNAs and the expression levels of these genes negatively correlated with siRNA levels. These *TAG*s included a cluster of *AGAMOUS-LIKE* genes (*AGL*s) and imprinted genes shown to play critical roles in endosperm development. We showed that these *AGL*s were subject to the regulation of p4-siRNAs.

Second, I tested the roles for maternal p4-siRNAs in RNA-directed DNA methylation (RdDM) in *Arabidopsis* endosperm. I found the majority of maternal p4-siRNAs were derived from short TEs in euchromatic loci and regulate RdDM and gene expression in spatiotemporal-specific manners. These euchromatic loci included *AGL91* and *AGL40*, which were actively expressed in chalazal endosperm where RdDM remained inactive up to the heart stage but silenced in other regions where RdDM is active. *AGL91* was paternally expressed, but *AGL40* was biparentally expressed. Maternal siRNAs mediated silencing of the maternal *AGL91* allele before fertilization and silencing of the paternal *AGL91* allele and *AGL40* after fertilization. Moreover, disrupting or overexpressing *AGL91* and *AGL40* altered seed size, providing evidence for *AGLs* in endosperm development and seed size control. This research uncovers a mechanism of gene regulation by maternal p4-siRNAs in endosperm and reveals their potential role in endosperm development and genomic imprinting.

Table of Contents

List of Tables xi

List of Figures xii

Chapter 1: Introduction 1

PolIV siRNAs 1

Biogenesis of p4-siRNAs 2

p4-siRNA generating loci 2

Molecular functions of p4-siRNAs 3

Maternal expression of p4-siRNAs in *Arabidopsis* endosperm. 5

Arabidopsis endosperm development 6

Embryogenesis and endosperm development in *Arabidopsis* 6

Parental genome dosage and endosperm development 7

Type I MADS box transcription factors 8

Epigenetic regulation in Endosperm 10

DNA demethylation 10

Trimethylation of histone 3 lysine 27 11

Genomic imprinting 12

Overview of dissertation research 14

Chapter 2: Maternal siRNAs As Regulators of Parental Genome Imbalance and Gene Expression in Endosperm of *Arabidopsis* Seeds 19

Abstract 19

Introduction 20

Results 22

Parent-of-origin effects on endosperm size and siRNA production in reciprocal triploids. 22

Ploidy-dependent siRNAs are derived from transposable elements (TEs) and TE-associated genes (TAGs) in endosperm. 23

Maternal siRNAs regulate *AGAMOUS-LIKE* (*AGL*) gene expression. 26

Expression of p4-siRNAs and *AGLs* in interploidy crosses is dependent on PolIV. 27
Discussion 28
Materials and Methods 30
Plant materials and growth conditions 30
Chromosome counts 30
Small RNA library construction 31
Bioinformatic analysis 31
Microarray analysis 32
Quantitative RT-PCR (qRT-PCR) analysis 32
Small RNA blot analysis 32
Chapter 3: Maternal siRNAs Guide Spatiotemporal Regulation of Euchromatic Loci and DNA Methylation in *Arabidopsis* Endosperm 52
Abstract 52
Introduction 52
Results 54
The parent-of-origin effect on euchromatic siRNA loci that are derived from short transposable elements (TEs) in endosperm 54
Euchromatic maternal siRNAs correlate with RdDM in seed 57
Spatial and temporal regulation of euchromatic siRNA genes and RdDM in endosperm 59
Effects for euchromatic siRNAs on seed size and embryogenesis 61
AGL91 and *AGL40* affect endosperm development and seed size 62
Discussion 62
Pol IV-associated siRNAs maintain genome stability in embryo and spatiotemporal regulation of euchromatic loci in endosperm 62
Expression of paternal *AGL91* allele in chalazal endosperm is regulated by maternal siRNAs before and after zygote formation 63
Roles for maternal siRNAs in imprinting and seed development in flowering plants 64
Materials and Methods 66
Plant materials and growth conditions 66

RNA extraction, cDNA synthesis and qRT-PCR analysis 66
Small RNA library preparation and sequencing 67
Analysis of small RNAs.....67
Plasmid construction.....68
GUS staining and microscopy 69
Cleared seed image 70
Seed size and weight measurement.....70
Analysis of DNA methylation using methylation-sensitive restriction enzyme PCR.....71

Chapter 4: Discussion and Future Directions 90
Biogenesis of p4-siRNAs from *AGL*s.....90
Transcriptional and post-transcriptional silencing.....91
Genomic imprinting of *AGL*s.....92
p4-siRNAs as species barrier 93

References.....96

Vita 110

List of Tables

Table 2.1: Sequence reads of small RNA libraries from the leaves of *Arabidopsis thaliana* reciprocal interploidy crosses and their parents 46

Table 2.2: Sequence reads of small RNA libraries from the seeds of *Arabidopsis thaliana* reciprocal interploidy crosses and their parents 46

Table 2.3: The expression changes (log2 fold change) of 27 validated genes that encode endosperm preferred early seed stage transcription factors..47

Table 2.4: A list of MADS box genes that possess more than 10 reads from transcribed regions in at least one line.....49

Table 2.5: 5' and 3' adaptor sequences for small RNA library construction.....49

Table 2.6: List of qPCR primers used for this study.....50

Table 2.7: Lists of probes used for this study. 51

Table 3.1: Summary of small RNA reads 87

Table 3.2: The primers for generating plasmid *pAGL91:AGL91::GUS:3'TE*...88

Table 3.3: The primers for generating plasmid *pAGL40:AGL40::GUS*.....88

Table 3.4: The primers for generating *AGL40* and *AGL91* overexpression lines.89

Table 3.5: qPCR primers for methylation sensitive restriction enzyme assay...89

List of Figures

Figure 1.1: Schematic diagram of PolIV-siRNA biogenesis and RNA-directed DNA methylation pathway in *Arabidopsis* 17

Figure 1.2: Endosperm development in *Arabidopsis* 18

Figure 2.1: Seed morphology and chromosome counts in interploidy crosses. ..34

Figure 2.2: Seed size variation in ploidy series (Col-0) and small RNA distribution in leaves. 35

Figure 2.3: Small RNA distribution in interploidy crosses. 36

Figure 2.4: miRNA and tasiRNA expression in ploidy series.....37

Figure 2.5: Distribution of TE-associated genes (*TAG*s) and parent-of-origin effects of siRNAs on *TAG* expression in reciprocal interploidy crosses.....38

Figure 2.6: Correlation of differentially expressed genes and siRNAs in interploidy crosses. 39

Figure 2.7: qRT-PCR analysis (REL, relative expression levels) of siRNA-related *AGL* genes in different stages of developing siliques. 40

Figure 2.8: Small RNA expression in the type I MADS box transcription factors and the gene expression patterns during female gametophyte and seed development. 41

Figure 2.10: siRNA profiles (Left) and qRT-PCR analysis (Right) of additional *AGL*s in 2 X 4 and 4 X 2 triploids and their parents (2 X 2 and 4 X 4).....43

Figure 2.11: siRNA production and *AGL* expression are dependent on RNA polymerase IV (NRPD1A) in endosperm and a model for endosperm development in interploidy crosses. 44

Figure 2.12: qRT-PCR (REL, relative expression levels) and small RNA blot analysis in interploidy crosses with *nrpd1a*.....45
Figure 3.1: Euchromatic siRNAs are derived from short TEs and mainly in endosperm. 72
Figure 3.2: Distribution of small RNAs, and chromosomal view of siRNA distribution. 73
Figure 3.3: Maternal dependence of euchromatic small RNA (sRNA) in developing seeds. 74
Figure 3.4: Distributions of siRNAs in F1 crosses involving *rdr2* and *nrpd1a* mutants. 75
Figure 3.5: Enrichment of 24-nt siRNAs in TE families of different size and distance to genes. 76
Figure 3.6: Association of maternal siRNAs with RNA-directed DNA methylation in endosperm. 77
Figure 3.7: Maternal siRNAs guide RNA-directed DNA methylation in endosperm. 78
Figure 3.8: Genes associated with maternally euchromatic siRNAs and endosperm DMRs. 79
Figure 3.9: Maternal siRNA loci are associated with RdDM in endosperm and embryo. 80
Figure 3.10: Spatiotemporal regulation of euchromatic maternal siRNA targets in endosperm. 81
Figure 3.11: Expression of paternal *AGL91* and activation of *AGL91* in RdDM mutants. 82

Figure 3.12: Enrichment of paternal biased genes in genes associated with maternal siRNA loci and endosperm DMRs. 83
Figure 3.13: Spatial and temporal regulation of *AGL40* by maternal siRNAs in endosperm. 84
Figure 3.14: Parent-of-origin effects of p4-siRNAs on seed size and embryogenesis. 85
Figure 3.15: Effects of *AGL* expression on seed size. 86

Chapter 1: Introduction

POLIV SIRNAS

Small RNAs are short RNA molecules that regulate gene expression and maintain genome stability in both animals and plants. Plants produce several different classes of small RNAs from endogenous genomic loci, including microRNA (miRNA), trans-acting small interfering RNA (tasiRNA), PolIV siRNA (p4-siRNA), and natural antisense siRNA (nat-siRNA) (Axtell, 2013; Baulcombe, 2004; Chapman and Carrington, 2007; Chen, 2009; Vazquez, 2006).

miRNA pathway is a posttranscriptional regulatory mechanism conserved in animal and plant kingdoms (Ambros, 2004; Bartel, 2009; Chen, 2009; Ruvkun, 2001). The majority of miRNAs in plant is 21-22 nt long and negatively regulates gene expression by cleaving the mRNAs with near-perfect match (Axtell et al., 2007; Bartel, 2009; Vazquez, 2006). miRNAs play pivotal roles in almost all aspects of plant development as well as responses to stimuli (Chen, 2012; Sunkar et al., 2012). Some non coding mRNAs targeted by miRNAs produce abundant secondary siRNAs around the cleavage sites in phased positions. These siRNAs act *in trans* to mediate the cleavage of mRNAs encoding other proteins and, thus, are denoted as tasiRNAs (Yoshikawa et al., 2005).

In addition to miRNAs and tasiRNAs, plants produce a plethora of siRNAs from transposable elements (TEs) and pericentromeric repeats. Biogenesis of siRNAs from transposons requires the function of RNA polymerase IV (Pol IV), and, thus, is denoted as Pol IV siRNA (p4-siRNA) (Mosher et al., 2008; Zhang et al., 2007). p4-siRNAs are usually 24-nt long and suppress transposon activities by mediating RNA-directed DNA methylation (RdDM) and chromatin modification (Chapman and Carrington, 2007; Herr et al., 2005; Onodera et al., 2005).

Biogenesis of p4-siRNAs

Biogenesis of p4-siRNAs depends on Pol IV, a plant specific polymerase (Herr et al., 2005). Pol IV initiates siRNA biogenesis by generating long single-stranded RNA transcripts preferentially from transposable elements (TEs) and heterochromatic sequences (Figure 1.1) (Pikaard et al., 2008; Wierzbicki et al., 2009). These transcripts are then converted by RNA-DEPENDENT RNA POLYMERASE 2 (RDR2) into double-stranded RNAs that are processed into 24-nt siRNAs by DICER-LIKE 3 (DCL3) and loaded into ARGONAUTE 4 (AGO4) (Herr et al., 2005; Kasschau et al., 2007; Lu et al., 2006; Xie et al., 2004). After Dicer cleavage from long dsRNA, p4-siRNAs are methylated at the 2' hydroxyl by the siRNA methyltransferase HUA ENHANCER 1 (HEN1) which prevents siRNA terminal uridylation and degradation (Li et al., 2005; Xie et al., 2004; Yu et al., 2005). Although *in vitro* transcriptional activity of PolIV has yet to be determined for PolIV, its *in vivo* function requires the Metal A and Metal B sites which catalyze transcription in other polymerases (Haag et al., 2009). The 5' and 3' end modifications of a PolIV transcript are also unknown, but might carry signal that shows the cell that the transcript is "aberrant" and thus recruit RDR2 (Mosher, 2011). Mutations in the largest subunit of PolIV, NUCLEAR RNA POLYMERASE D1 (NRPD1), including mutations within the conserved metal binding motif, deplete the expression of 24-nt siRNAs. The SWI/SNF nucleosome remodeling protein CLASSY1 (CLSY1) is required for p4-siRNA biogenesis at many genomic loci probably by recruiting PolIV to chromatin (Smith et al., 2007).

p4-siRNA generating loci

The advent of next-generation sequencing technology has allowed researchers to get the sequences of millions of small RNA molecules. In *Arabidopsis*, the vast majority of unique small RNAs are 24-nt p4-siRNAs, many of which are present at very low

frequency (<1 per million) (Henderson et al., 2006; Nobuta et al., 2008). Unlike miRNAs, which are usually cleaved from a defined position in the hairpin-structured precursor, p4-siRNAs are cleaved at random from dsRNA precursor. Therefore, a 500-bp genomic fragment has the potential to generate almost 1000 unique p4-siRNA sequences (Mosher, 2011). p4-siRNAs are produced from thousands of distinct genomic loci, covering a minimum of 1% of the *Arabidopsis* genome and a much higher fraction of maize and wheat genomes which are abundant in TEs (Cantu et al.; Mosher, 2011; Mosher et al., 2008; Nobuta et al., 2008). Given the high sequence complexity of p4-siRNA population, it is necessary to consider p4-siRNA loci which are clusters of p4-siRNA sequences related to each other when mapped to the genome, rather than individual p4-siRNA sequences (Lu et al., 2006; Mosher, 2011; Mosher et al., 2008; Zhang et al., 2007).

p4-siRNAs usually match repetitive sequences including tandem and dispersed repeats from pericentromeric heterochromatin, rDNA arrays, and TEs (Cantu et al.; Kasschau et al., 2007; Mosher et al., 2008; Nobuta et al., 2008; Zhang et al., 2007). p4-siRNAs are also produced from unique genomic regions including intergenic sequences and protein coding genes that are not present elsewhere in the genome (Chan et al., 2006; Mosher et al., 2008; Xie et al., 2004).

Molecular functions of p4-siRNAs

p4-siRNAs induce *de novo* DNA methylation at asymmetric cytosines and are involved in histone modification and active DNA demethylation.

DNA methylation can occur at three sequence contexts in plants: CG, CHG and CHH (H= A, T or C), where CG and CHG are symmetric methylation while CHH is asymmetric methylation. Symmetric CG and CHG methylation is responsible for

maintenance methylation after each cell division. DNA METHYLTRANSFERASE 1 (MET1), the plant homolog of Dnmt1, is required for CG methylation, whereas CHG methylation depends on CHROMOMETHYLASE (CMT3) (Law and Jacobsen, 2010). By contrast, asymmetric CHH methylation is a hallmark of *de novo* methylation because it must be reestablished after each cell division. p4-siRNAs derived from methylated DNA before replication can guide *de novo* methylation to faithfully reproduce DNA methylation patterns after replication (Law and Jacobsen, 2010), a process known as RNA-directed DNA methylation (RdDM) (Figure 1.1). After being generated from repetitive sequences by PolIV, p4-siRNAs will bind to AGO4 and base pair with nascent RNAs transcribed by PolV, which is another plant specific polymerase that transcribes scaffold RNAs for RdDM (Wierzbicki et al., 2008; Wierzbicki et al., 2009). Consequently, AGO4 recruits the *de novo* methyltransferase DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) to methylate the cytosines in CHH context (Law and Jacobsen, 2010).

Changes in histone methylation are also the direct or indirect consequence of p4-siRNA guided RdDM. Histone H3 lysine 9 dimethylation (H3K9me2) and lysine 27 monomethylation (H3K27me1), the hallmarks for silent heterochromatin, are reduced at some p4-siRNA loci when biosynthesis of p4-siRNAs is disrupted (Huettel et al., 2006; Wierzbicki et al., 2008; Wierzbicki et al., 2009; Xie et al., 2004; Zilberman et al., 2003). Cytological changes in heterochromatin are also detected in PolIV mutants including disassociation of nucleolar organizing regions (NORs), decondensation of chromocenters, and dispersal of heterochromatic foci (Onodera et al., 2005).

In addition to DNA methylation and histone methylation, there is evidence that p4-siRNAs are associated with active DNA demethylation. In *Arabidopsis*, REPRESSOR OF SILENCING 1 (ROS1) and its homologs DEMETER-LIKE 2 (DML2) and

DEMETER-LIKE 3 (DML3) act as DNA glycosylase/demethylase to excise 5-methyl cytosine (Huettel et al., 2006; Penterman et al., 2007). ROS1 activity might be upregulated by the PolIV/V pathway to generate negative feedback and inhibit RdDM activity (Penterman et al., 2007). It is probably through the physical interaction with REPRESSOR OF SILENCING 3 (ROS3), which binds 24-nt siRNAs, that ROS1 family proteins are targeted to specific genomic loci (Zheng et al., 2008).

Maternal expression of p4-siRNAs in *Arabidopsis* endosperm.

Interestingly, p4-siRNAs are expressed only from maternal alleles in *Arabidopsis* endosperm (Mosher et al., 2009). By contrast, tasiRNAs are expressed from both chromosomes. During vegetative growth, both alleles from maternal and paternal origins are detectable, indicating that imprinting of p4-siRNAs is limited to endosperm development. Since p4-siRNA levels rise dramatically after fertilization, maternal p4-siRNAs in endosperm cannot be a result of maternal carry-over, but rather must result from genomic imprinting (Mosher, 2011; Mosher et al., 2009). Accumulation of maternal p4-siRNAs in endosperm requires biogenesis of p4-siRNAs in the female gametophyte, suggesting that p4-siRNA production from maternal chromosomes before fertilization provides feedback to recruit PolIV to the same chromosome after fertilization (Mosher, 2011; Mosher et al., 2009). A burst of p4-siRNA expression was detected at 4 to 6 days after pollination (DAP) in developing endosperm, but not in embryo, which is reminiscent of the enrichment of 21-nt siRNAs in male gametes (Slotkin et al., 2009). In *Arabidopsis* pollen, TEs are unexpectedly reactivated and transpose only in the pollen vegetative nucleus (VN), which accompanies the sperm cells but does not transmit DNA to the fertilized zygote. VN may contribute siRNAs to sperm cells and reprogram the transposon silencing in next generation (Slotkin et al., 2009). As companion cells that

nourishes developing embryo, endosperm may also produce siRNAs to suppress transposon reactivation in embryo (Ibarra et al., 2012; Mosher and Melnyk, 2010).

ARABIDOPSIS ENDOSPERM DEVELOPMENT

Endosperm is the nutritive tissue surrounding embryo and provides nutrients to support embryo growth. Seeds in higher plants are formed by a unique process in flowering plants known as "double fertilization" (Berger, 2008). During fertilization, the pollen tube delivers two sperms to the embryo sac. One sperm fuses with the egg cell to give rise to the embryo and the other sperm fuses with the central cell to generate endosperm. Since the central cell contains two haploid nuclei, the endosperm resulted from the second fertilization is triploid which contains two maternal nuclei and one paternal nuclei.

In addition to its biological function, endosperm stores carbohydrates and proteins during and represents the major source of food, feed and fuel for humans. It is estimated that crop seeds provide nearly 70-80% of calories and 60-70% of all proteins consumed by the human population (Borlaug, 1973).

Embryogenesis and endosperm development in *Arabidopsis*

Upon double fertilization, *Arabidopsis* seed development undergoes three major events: (1) Postfertilization to proembryo (preglobular) stage when terminal and basal cell differentiate to form suspensor and embryo proper; (2) Globular to heart transition stage when major tissue-type primordia are differentiated and radial axis is established. At this stage, embryo proper becomes bilaterally symmetrical and shoot-root axis becomes visible; (3) Organ expansion and maturation stage when cotyledons enlarge and storage proteins and lipids accumulate followed by dehydration and dormancy (Goldberg et al., 1994).

Endosperm development in *Arabidopsis* is initiated by repeated divisions of the triploid nucleus without cytokinesis. The multinucleate cell during initial stages of development is known as syncytium, which is an unusual structure in higher plants (Brown et al., 1999). Endosperm stages are defined by successions of pseudo-synchronous mitoses (Berger, 2003) (Figure 1.2). The nuclei then migrate to develop three distinct regions: the region surrounding the embryo (MCE), the central or peripheral endosperm (PEN), and the region of the chalazal endosperm (CZE) (Olsen, 2004) (Figure 1.2). Seed enlarges rapidly during syncytium stage and is occupied mostly by the endosperm. Partition of the multinucleate cell into individual mononucleate cell, termed cellularization, occurs after the eighth mitotic division when the embryo reaches the heart stage except the nuclei at CZE (Berger, 2003). In this subregion, nuclei are grouped in cytoplasmic pockets and cellularization is believed to eventually take place at a later stage (Brown et al., 1999). After cellularization, cell divisions in endosperm are rare whereas embryonic cells divide rapidly and absorb the endosperm. Upon seed maturation, the endosperm dies and the embryo eventually occupies most of the seed (Berger, 1999).

Parental genome dosage and endosperm development

Normal seed development requires an "optimum" maternal to paternal genome ratio of 2:1 (2m:1p) in the triploid endosperm nucleus. The timing of cellularization, seed size, and seed viability are greatly affected by this ratio, which has been demonstrated in the interploidy hybrids. In *Arabidopsis thaliana*, increasing the paternal genome ratio (2m:2p or 2m:3p) in endosperm by pollinating a diploid "mother" with a tetraploid or hexaploid "father" (2X4 or 2X6) produced larger or collapsed seeds. By contrast, increasing the maternal genome ratio (4m:1p or 6m:1p) in endosperm by pollinating a

tetraploid or hexaploid mother with a diploid father (4X2 or 6X2) produced smaller seeds (Scott et al., 1998; Tiwari et al., 2010). The difference in seed size resulted from imbalanced parental dosage is a result of altered timing of endosperm cellularization (EC). The timing of EC is positively correlated with endosperm proliferation and seed size but negatively correlated with increased maternal genome dosage (Figure 1.2). The observation that an increased maternal genome dosage leads to smaller seed size and an excessive dosage of paternal genome results in bigger seed size, is in agreement with the predictions of the "parental conflict theory" (Moore and Haig, 1991) or "kinship theory" (Trivers and Burt, 1999), which postulate that increased maternal genome dosage restricts nutrient flow to the embryo whereas increased paternal genome dosage enhances nutrient flows to the embryo (Haig and Westoby, 1989). Emerging evidence has revealed that genomic imprinting, a phenomenon by which certain genes are expressed in a parent-of-origin manner, might be the mechanism underlying parental genome dosage balance in endosperm development (Erilova et al., 2009; Jullien and Berger, 2010; Tiwari et al., 2010).

Type I MADS box transcription factors

One of the largest groups of transcription factors enriched in developing seeds are Type I MADS-box transcription factors (Bemer et al., 2010). Type I genes can be further subdivided into three groups: M_{α}, M_{β}, and M_{γ} classes based on the sequence of the MADS box and presence of additional motifs (De Bodt et al., 2003a; Parenicova et al., 2003). Many of M_{α} class and the majority of M_{γ} class genes are highly expressed in endosperm, some of which have established functions in central cell and endosperm development (Bemer et al., 2010). *AGAMOUS-LIKE80* (*AGL80*) (M_{γ} class), and *AGL61* (*DIA*; M_{α} class), are proposed to form a complex to specify the formation of the central

cell in the embryo sac. Mutations in both *AGL80* and *AGL61* impaire central cell development, leading to a maternal-lethal phenotype (Bemer et al., 2008; Portereiko et al., 2006; Steffen et al., 2008). *AGL62* (M_α class) and *AGL37* (*PHE1*; M_γ class) are involved in endosperm development. *AGL62* controls the timing of endosperm cellularization. In *agl62* mutants, the endosperm undergoes precocious cellularization, resulting in arrest of embryo growth (Kang et al., 2008). *AGL37* is regulated by MEA, a component of PRC2, after fertilization and is paternally expressed. Although *agl37* mutants do not show abnormal endosperm phenotype, knocking down expression levels of *AGL37* in *mea* seeds partially ameliorate the mutant phenotype (Kohler et al., 2003). Moreover, *AGL36* (M_γ class) is maternally expressed in endosperm and *agl36* mutants display wild-type phenotype, suggesting sequence redundancy in Type I subfamily (Shirzadi et al., 2011). There are 61 Type I genes in *Arabidopsis* genome and none of them are discovered in forward genetic studies (Bemer et al., 2010). Type I subfamily has undergone many recent duplications and these genes are subject to high birth and death rates, suggesting several type I genes may undergo nonfunctionalization (Bemer et al., 2010; De Bodt et al., 2003b; Nam et al., 2004).

Interestingly, a cluster of Type I genes (M_α and M_γ classes) are upregulated in the seeds resulting from interploidy hybrids and interspecific hybrids in *Arabidopsis* where endosperm undergoes delayed cellularization and overproliferation (Tiwari et al., 2010; Walia et al., 2009). Endosperm failure eventually leads to arrest of embryo growth and hence seed abortion (Bushell et al., 2003). Reduced expression of some *AGL*s ameliorates seed abortion phenotype in *Arabidopsis* interspecific hybrids (Walia et al., 2009).

EPIGENETIC REGULATION IN ENDOSPERM

Central cell and endosperm undergoes substantial epigenetic reprogramming that are essential for normal endosperm development and seed viability. The best characterized processes are DNA demethylation and PcG-mediated histone methylation. These epigenetic regulations are part of the mechanism underlying the genomic imprinting which specifically occurs in the endosperm.

DNA demethylation

Endosperm undergoes extensive DNA demethylation at all sequence contexts in *Arabidopsis* (Gehring et al., 2009; Hsieh et al., 2009). Mutations in DNA demethylase DEMETER (DME) produce aborted seeds, suggesting that DNA demthylation is required for endosperm development (Choi et al., 2002). DME excises 5' methylated cytosines in the central cell and results in demethylated maternal chromosomes in the endosperm (Gehring et al., 2009; Hsieh et al., 2009). DNA demethylation in the endosperm depends not only on DME activity, but also on repression of MET1 in central cell before fertilization as well as in endosperm after fertilization (Hsieh et al., 2011; Jullien et al., 2008). Repression of *MET1* in central cell has been shown to depend on the retinoblastoma pathway (Jullien et al., 2008), whereas maintenance of *MET1* silencing in the central cell and endosperm likely depends on PRC2 activity (Hsieh et al., 2011). Therefore, it seems plausible that DME is required for demethylation of a subset of loci, whereas global demethylation is at least in part a consequence of *MET1* repression in the central cell (Hsieh et al., 2011).

Global reduction of DNA methylation occurs across the endosperm genome compared to embryo, preferably at TEs and regions enriched in p4-siRNAs (Hsieh et al., 2009). Interestingly, most known imprinted genes are associated with differential methylated regions (DMR) which are hypomethylated in endosperm compared to

embryo. It is predicted that there were approximately 50 imprinted genes in *Arabidopsis* genome based on the presence of DMRs and enrichment of mRNAs in endosperm (Gehring et al., 2009). Therefore, demethylation in endosperm seems to be necessary for parent-of-origin expression of genes. Indeed, *DME* is required for the maternal expression of *FWA* and *MEA* after fertilization, suggesting that removal of DNA methylation from maternal chromosome is necessary for the activation of maternal allele (Gehring et al., 2006; Kinoshita et al., 2004).

DNA demethylation also occurs in the companion cells of gametes. The male gametophyte contains two sperm cells and a vegetative cell which serves as nursing cell. The vegetative nucleus undergoes demethylation and reinforces CHH methylation and TE silencing in sperms. In *dme* mutant pollens, vegetative nucleus is hypermethylated and CHH methylation is reduced in sperm (Calarco et al., 2012; Ibarra et al., 2012; Slotkin et al., 2009). As the companion cells for embryo, similar mechanism has been proposed for endosperm that it sacrifices its own genome integrity to protect the embryo from reactivation and transposition of TEs (Ibarra et al., 2012; Mosher et al., 2011). Although attractive, it is yet to be examined whether CHH methylation is indeed reduced and TEs are reactivated in the embryo of *dme* mutants.

Trimethylation of histone 3 lysine 27

Components of Polycomb repressive complex 2 (PRC2) have been identified in forward genetic studies for their critical functions during endosperm development (Hennig and Derkacheva, 2009). PcG proteins are evolutionarily conserved master regulators of cell identity and coordinate decisions between cell proliferation and cell differentiation (Schuettengruber and Cavalli, 2009). PRC2 catalyzes the trimethylation of histone H3 on lysine 27 (H3K27me3) during vegetative and reproductive plant

development (Schuettengruber and Cavalli, 2009). The FIS PRC2 complex contains four subunits: MEA, FIS2, FERTILIZATION INDEPENDENT ENDOSPERM (FIE), and MULTICOPY SUPPRESSOR OF IRA1 (MSI1). The FIS complex is active in central cell and endosperm before and after fertilization and plays essential roles in endosperm development (Hennig and Derkacheva, 2009). Mutations in components of FIS complex cause endosperm overproliferation and seed abortion (Leroy et al., 2007; Spillane et al., 2000).

Direct target genes of the FIS complex include PHE1 (Kohler et al., 2003), FUSCA3 (Makarevich et al., 2006) and MEA itself (Baroux et al., 2006; Gehring et al., 2006; Jullien et al., 2006). *MEA* is expressed from the maternal allele and represses the paternal expression to maintain the self imprinting of *MEA* after fertilization (Baroux et al., 2006; Gehring et al., 2006; Jullien et al., 2006). Genome-wide profiling of H3K27me3 in endosperm has identified target genes that have functional roles in endosperm cellularization and chromatin architecture (Weinhofer et al., 2010). A subset of transposable elements (TEs) that are protected by DNA methylation in vegetative tissues are targeted by the FIS complex in endosperm, suggesting that DNA methylation and H3K27me3 compensate for each other in repressing a subset of TEs (Weinhofer et al., 2010).

Genomic imprinting

Genomic imprinting is an epigenetic phenomenon found in flowering plants and animals whereby certain genes are expressed in a parent-of-origin-specific manner. In flowering plants, gene imprinting occurs only in endosperm and is rarely detected in other tissues (Costa et al., 2012; Huh et al., 2007; Pignatta and Gehring, 2012). Until recently, only a handful of imprinted genes were known in plants. High-throughput

sequencing has greatly advanced the discovery of imprinted genes. Studies using deep sequencing of mRNA libraries derived from reciprocal intraspecific crosses have confirmed that imprinting is mostly endosperm-specific and have significantly expanded the number of imprinted genes to about 50 in *Arabidopsis thaliana* (Gehring et al., 2011; Hsieh et al., 2011; Wolff et al., 2011). Since most imprinted genes are identified recently, only 4 MEGs are well characterized for their functions in development (Raissig et al., 2011). The two Polycomb group genes *MEA* and *FIS2* inhibit endosperm development without fertilization, and mutation in either of the two genes display autonomous endosperm proliferation with and without fertilization (Chaudhury et al., 1997; Grossniklaus et al., 1998; Ingouff et al., 2005b; Kiyosue et al., 1999). The embryogenesis of the mutants is delayed and the mutant embryos only reach the late heart or sometimes torpedo stage (Chaudhury et al., 1997; Grossniklaus et al., 1998; Ingouff et al., 2005b; Kiyosue et al., 1999). Similarly, mutation in *ARABIDOPSIS FORMIN HOMOLOGUE 5* (*FH5*) and *MATERNALLY EXPRESSED PAB C-TERMINAL* (*MPC*) display defects in endosperm development (Fitz Gerald et al., 2009; Ingouff et al., 2005a; Tiwari et al., 2008). Therefore, the best characterized imprinted genes all share a role in endosperm development or are at least preferentially expressed in this tissue (Raissig et al., 2011).

Some of the imprinted genes are deregulated when parental genome dosage is altered. One of the MEGs, *MEA*, is found to be downregulated in response to doubled paternal genome. *MEA* is one of the components of FIS PRC2 complex and hence negatively regulates the expression of many transcription factors. It is proposed that *MEA* senses the changes in parental genome dosage and the downregulation of *MEA* leads to the upregulation of many downstream genes in paternal access seeds (Erilova et al., 2009). Moreover, imprinting patterns of *PHE1* (PEG), and *MEA* (MEG) are lost in interploidy hybrids (Jullien and Berger, 2010). Deregulation of imprinted genes is likely

to be the consequence of the disrupted balance of epigenetic regulation in parental genomes. Indeed, hypomethylated maternal or paternal genome resulted from reciprocal crosses between DNA methylation mutants and wild type plants phenocopies interploidy hybrids (Adams et al., 2000).

Although only a limited number of protein coding genes are imprinted in *Arabidopsis*, study on non-coding RNAs reveals thousands of maternally imprinted p4-siRNA loci during endosperm development (Mosher et al., 2009). Interestingly, imprinted genes are frequently associated with TEs or internal repeats and in some cases these repetitive sequences are sufficient to recapitulate the imprinting pattern of the genes (Kinoshita et al., 2007; Villar et al., 2009). Since p4-siRNAs are preferentially generated from TEs, it is possible that the imprinting of protein coding genes is the byproduct of p4-siRNA imprinting.

OVERVIEW OF DISSERTATION RESEARCH

p4-siRNAs represent the largest group of imprinted genomic loci in endosperm. Endosperm development is accompanied by a burst of p4-siRNAs at 4-6 days after pollination, coincident with cellularization of the syncytium (Mosher et al., 2009). However, the molecular function of p4-siRNAs in seed development remains enigmatic. It is estimated that 70-100% more food will need to be produced worldwide by 2050 without an appreciable increase in arable land and despite global climate change (Godfray et al., 2010). A detailed understanding of seed development will provide cogent targets and strategies to improve seed quality and yield. This dissertation research aims to unravel the role of p4-siRNAs in endosperm during early seed development in *Arabidopsis*.

The dissertation is divided into two related subjects. First, I generated and analyzed the genome-wide small RNA profiles in developing seeds of *Arabidopsis*. Interploidy crosses disrupt parental genome dosage in endosperm and alter seed size. I hypothesized p4-siRNAs are the maternal genome-specific factors that respond to the imbalance of parental genome dosage and contribute to gene expression changes. To test this hypothesis, I analyzed the small RNA profiles using high throughput sequencing as well as existing transcriptome data in the seeds of reciprocal interploidy crosses and their parents. Levels of 24-nt p4-siRNAs were proportional to maternal genome dosage and negatively correlated with protein-coding genes associated with TEs (*TAG*s), which generated considerable amount of p4-siRNAs. To test the role of p4-siRNAs in seed development, I identified *TAG*s that were specifically expressed in endosperm, which included a group of Type I MADS-box transcription factors. I examined the maternal siRNA levels as well as *AGL* expression levels in *nrpd1a* mutant. These experiments will be described in Chapter 2. This work established the link of maternal siNRAs to gene expression in *Arabidopsis* seeds and provided candidate genes for further study.

Second, following the previous findings, I proposed the molecular mechanism of p4-siRNA-mediated gene silencing in endosperm. First, I identified genomic loci that were prone to maternal p4-siRNA accumulation by high throughput small RNA sequencing in PolIV pathway mutants. I found maternal siRNAs are preferably generated from euchromatic loci enriched in short TEs and protein-coding genes. By further dissecting the seed into endosperm and embryo, I found that euchromatic maternal siRNAs were mainly produced in endosperm rather than embryo. To test how p4-siRNAs affect DNA methylation and gene expression in seed, I analyzed existing whole-genome bisulfite sequencing data and tissue-specific transcriptome data in seeds. I found euchromatic maternal siRNAs guide DNA methylation to their cognate loci and these loci

were subject to spatiotemporal regulation by RdDM. Using stable transgenic plants, I examined spatial and temporal expression of *AGL91* and *AGL40* in both wild type and PolIV mutants. In the absence of p4-siRNAs, *AGL* expression was reactivated in the subregions that did not express *AGL* in wild type and was prolonged to a later developmental stage. To determine the effects of maternal p4-siRNAs and *AGL*s on seed size and development, I examined seed size and embryogenesis process in PolIV mutants and *AGL* mutants. I also generated transgenic plants that overexpressed *AGL40* and *AGL91*. I found PolIV inhibits endosperm growth while *AGL40* and *AGL91* positively affected seed size. These experiments will be described in Chapter 3. This work elucidated the molecular mechanism of p4-siRNA-mediated gene silencing in endosperm and revealed the spatiotemporal regulation of RdDM in *Arabidopsis* endosperm.

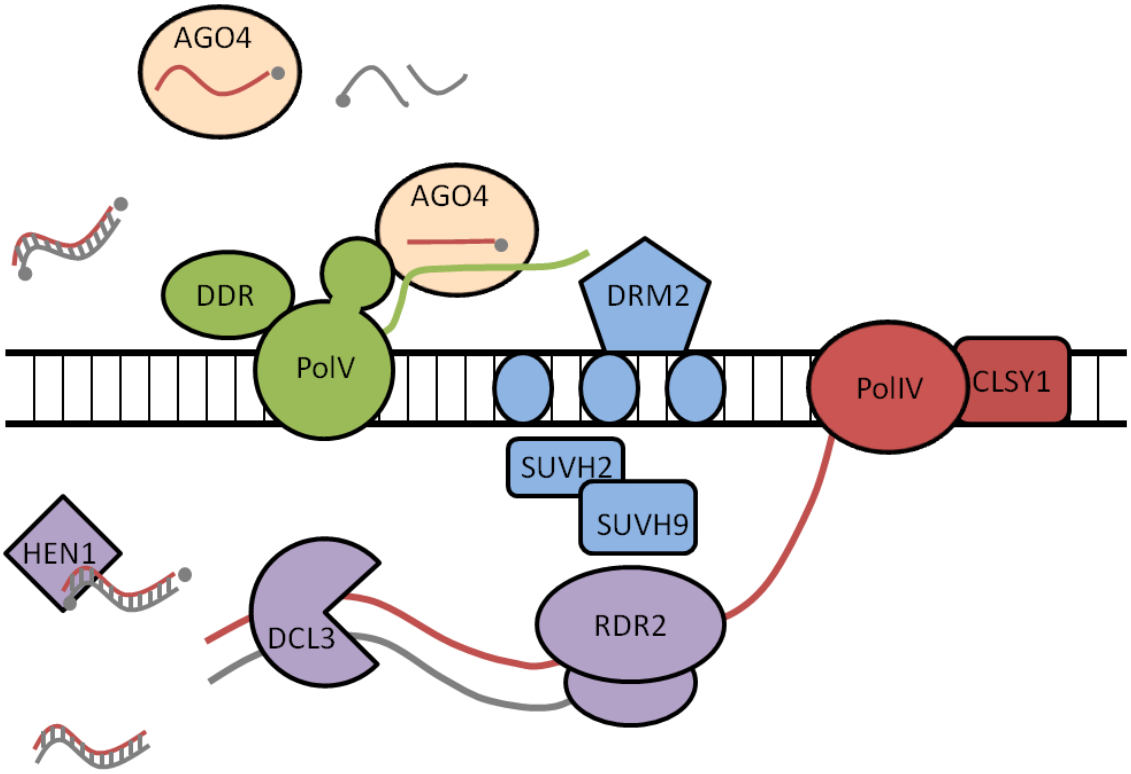

Figure 1.1: Schematic diagram of PolIV-siRNA biogenesis and RNA-directed DNA methylation pathway in *Arabidopsis* (Adapted from (Mosher, 2011)). RNA polymerase IV (PolIV, red oval) transcribes p4-siRNAs from repetitive sequences with the assistance of the SWI/SNF nucleosome remodeling protein CLSY1 (red rectangle). RNA-dependent RNA polymerase 2 (RDR2, purple ovals) converts PolIV transcripts into double-stranded (ds)RNA which is subsequently cleaved into 24-nt siRNAs by Dicer-like ribonuclease 2 (DCL3, purple pac-man). The 24-nt siRNA duplexes are then methylated by HUA ENHANCER 1 (HEN1, purple diamond). ARGONAUTE 4 (AGO4, yellow oval) binds the siRNA duplexes and degrades one strand, leaving one p4-siRNA loaded in the protein. RNA Polymerase V (Pol V, green joined circles) transcribes scaffold RNAs for p4-siRNA binding with the assistance of DDR comples (green oval). AGO4 physically interacts with the carboxyl-terminal domain of Pol V and the p4-siRNA anneals to the nascent PolV transcripts. AGO4 recruits DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2, blue pentagon) to catalyze CHH (H = A, T or C) methylation (blue circles). Su (var)3-9 homologs SUVH2 and SUVH9 (blue rectangles) are also required for DNA methylation.

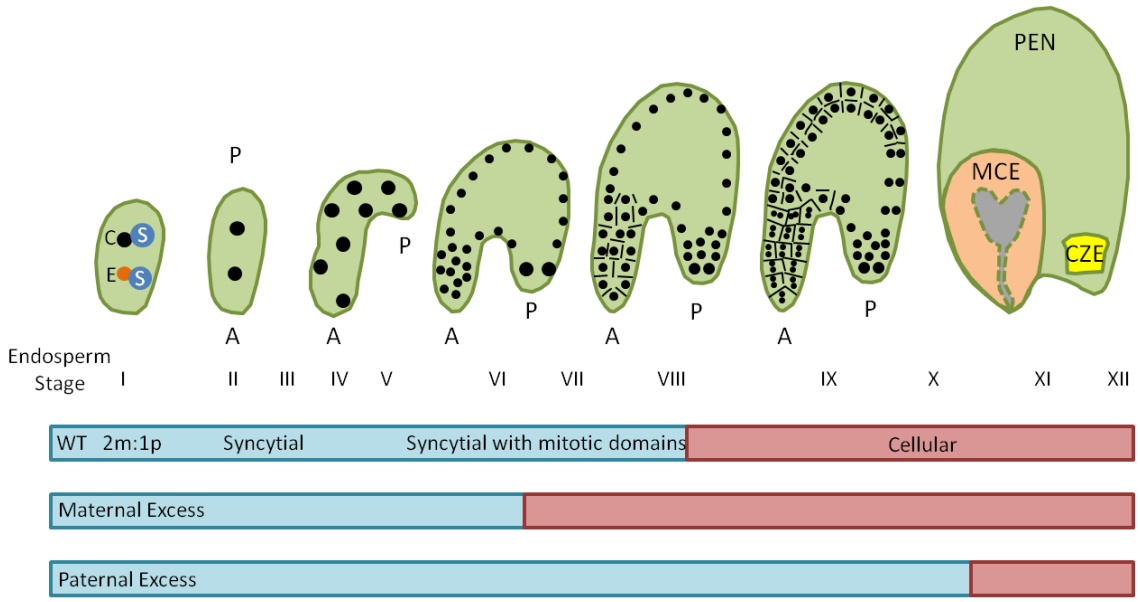

Figure 1.2: Endosperm development in *Arabidopsis* (Adapted from (Berger, 2003)). Upon double fertilization, one sperm cell (S, blue circle) fuses with egg cell (E, orange circle) to give rise to embryo while another sperm cell fuses with central cell (C) to form endosperm. Endosperm development can be divided in two major phases, first the syncytial and then the cellular phase. Endosperm stages are defined by successions of pseudo-synchronous mitoses. Three domains are defined from the anterior pole (A) to the posterior pole (P): mycropylar endosperm (MCE, orange), pheripheral endosperm (PEN, green) and chalazal endosperm (CZE, yellow). The eighth mitotic cycle is followed by cellularization of the syncytial peripheral endosperm and chalazal endosperm remains syncytial until an undefined late stage. Maternal or paternal excess results in precocious or delayed cellularization of syncytial endosperm.

Chapter 2: Maternal siRNAs As Regulators of Parental Genome Imbalance and Gene Expression in Endosperm of *Arabidopsis* Seeds[1]

ABSTRACT

Seed size is important to crop domestication and natural selection and is affected by the balance of maternal and paternal genomes in endosperm. Endosperm, like placenta in mammals, provides reserves to the developing embryo. Interploidy crosses disrupt the genome balance in endosperm and alter seed size. Specifically, paternal-excess crosses (2X4) delay endosperm cellularization (EC) and produce larger seeds, whereas maternal-excess crosses (4X2) promote precocious EC and produce smaller seeds. The mechanisms for responding to the parental genome dosage imbalance and for gene expression changes in endosperm are unknown. In plants, RNA polymerase IV (PolIV or p4) encoded by *NRPD1a* is required for biogenesis of a major class of 24-nt siRNAs (p4-siRNAs), which are predominately expressed in developing endosperm. Here we show that p4-siRNA accumulation depends on the maternal genome dosage, and maternal p4-siRNAs target transposable elements (TEs) and TE-associated genes (*TAGs*) in seeds. The p4-siRNAs correlate negatively with expression levels of AGAMOUS-LIKE (*AGL*) genes in endosperm of interploidy crosses. Moreover, disruption of maternal *NRPD1a* expression is associated with p4-siRNA reduction and *AGL* up-regulation in endosperm of reciprocal crosses. This is the first genetic evidence for maternal siRNAs in response to parental genome imbalance and in control of transposons and gene expression during endosperm development.

[1] This chapter is reformatted from "Maternal siRNAs as regulators of parental genome imbalance and gene expression in endosperm of Arabidopsis seeds." by Lu, J., Zhang, C., Baulcombe, D., and Chen, Z. J. (2012). Proc. Natl. Acad. Sci. USA vol. 98, pp. 200-205.

INTRODUCTION

Crop seeds provide nearly 70-80% of calories and 60-70% of all proteins consumed by the human population (Borlaug, 1973). Endosperm is the direct or indirect source for most of the nutritional content of the seed, and it is similar to the placenta in mammals (Moore and Haig, 1991), which is the source of nutrition for embryo development (Stebbins, 1976).

In angiosperms the endosperm is formed after pollination of the egg by a male gamete (pollen) that contains two sperm nuclei. One sperm fertilizes the egg to form a zygote with a 1:1 maternal to paternal genome ratio (1m: 1p), whereas the other fertilizes two central cell nuclei to form an endosperm cell with a 2:1 maternal to paternal genome ratio (2m:1p). In *Arabidopsis thaliana*, increasing the paternal genome ratio (2m:2p) in endosperm by pollinating a diploid "mother" with a tetraploid "father" (2X4) delays endosperm cellularization (EC) and produces larger seeds. By contrast, increasing the maternal genome ratio (4m: 1p) in endosperm by pollinating a tetraploid mother with a diploid father (4X2) leads to precocious EC and smaller seeds (Scott et al., 1998; Tiwari et al., 2010).

Transcription factors including AGAMOUS-LIKE proteins (AGLs) affect endosperm development (Bemer et al., 2010; Kang et al., 2008; Parenicova et al., 2003). AGLs are members of the plant type I MADS domain subfamily (Parenicova et al., 2003), and it is likely that they have a role in reproductive development because they are expressed in female gametophyte or developing seeds (Bemer et al., 2010). Mutations in *AGL62* lead to precocious EC and arrest of embryo growth, suggesting a direct effect of *AGL62* in endosperm development (Kang et al., 2008). *AGL36* is maternally imprinted and has a potential role in endosperm development, although no obvious phenotype is found in the *agl36* mutant probably because of redundancy in this subfamily (Shirzadi et

al., 2011). Moreover, up-regulation of *AGL62* and *AGL90* is related to the postzygotic barrier between *A. thaliana* and *A. arenosa*, which is associated with endosperm overproliferation and delayed development (Josefsson et al., 2006), similar to that in paternal-excess interploidy crosses.

Mechanisms for responding to parental genome dosage and for regulating *AGL* expression in endosperm are largely unknown. The model of parental genome balance to explain this effect requires a parent-of-origin-specific factor and a mechanism for balancing the level of this factor relative to the other parental genome. In principle this parent-of-origin-specific factor could involve imprinted genes, including *MEA* in *A. thaliana* (Erilova et al., 2009) and *PEG1* and *FIE101* in maize (Gutierrez-Marcos et al., 2003). However, expression patterns of many imprinted genes are contradictory to the predictions in interploidy crosses (Jullien and Berger, 2010; Tiwari et al., 2010). For example, maternally expressed genes including *MEA*, *FWA* and *FIS2* are up-regulated in the paternal-excess endosperm, similar to that of paternally expressed genes such as *PHE1*. It is therefore unlikely that they are the parent-of-origin-specific factor.

An alternative mechanism could involve 24-nt small interfering (si)RNAs that are dependent on *NRPD1a*. The NRPD1a protein is the largest subunit of RNA polymerase IV (PolIV or p-4), a homolog of DNA-dependent RNA polymerase II (Herr et al., 2005; Mosher et al., 2009; Onodera et al., 2005), and p4-siRNAs in developing seed of *A. thaliana* are predominately expressed from the maternal genome in endosperm (Mosher et al., 2009). Some 24-nt siRNAs are associated with target genes in leaves of *A. thaliana* hybrids (Groszmann et al., 2011). Here we test the possibility that maternally expressed p4-siRNAs are the factors sensitive to the parental genome dosage and regulate *AGL* expression levels in seeds through a mechanism in which, as in leaves of *A. thaliana* hybrids (Groszmann et al., 2011), they silence gene expression.

RESULTS

Parent-of-origin effects on endosperm size and siRNA production in reciprocal triploids.

In *A. thaliana*, reciprocal interploidy crosses between diploid (2X) and tetraploid (4X) plants produce variable seed sizes in Col-0 or C24 ecotypes (Figures 2.1A, 2.1C, and 2.2A-C). These plants contain the expected ploidy number of chromosomes (Figure 2.1B). As reported previously (Scott et al., 1998), the excessive paternal-genome dosage (2 X 4) results in larger seeds, whereas the excessive maternal-genome dosage (4 X 2) leads to smaller seeds. The response to paternal genome excess in 2 X 4 crosses was dependent on genotypes. Larger and normal seeds were produced in L*er* and C24, whereas in Col the seeds were aborted (Figure 2.2A) during seed coat development because the expression of *TTG2* and other genes was disrupted (Dilkes et al., 2008). However, during early seed development the response to altered parental genome dosage is consistent among all ecotypes tested. The endosperm size, reflected in the seed size (Scott et al., 1998), was noticeably different in the paternal- and maternal-excess seeds 5-6 days after pollination (DAP) in Col-0 (Figure 2.1D).

We manually dissected hundreds of seeds (containing endosperm and embryo) out of siliques at 6 DAP of reciprocal crosses (2 X 4 and 4 X 2) and their Col-0 parents, diploids (2 X 2) and tetraploids (4 X 4). At this stage, expression of p4-siRNA is most abundant in endosperm (Mosher et al., 2009), and the seed size was obviously different between reciprocal crosses (Figure 2.1D). Eight small RNA libraries were made from these immature seeds and from rosette leaves in reciprocal crosses and their parents. A total of ~80-million small RNA reads were generated by Illumina sequencing and ~ 64-million reads (~80%) were mapped (Tables 2.1 and 2.2). To reduce ambiguity, only the reads that perfectly matched sequences of the annotated genome (TAIR9) were

normalized to reads per 10 millions for further analysis. In seeds, the most abundant small RNAs were 21- and 24-nt long, representing 20-29% and 37-53% of total small RNAs (Figure 2.3A). The proportions of 21-nt (18-23%) and 24-nt (46-50%) small RNAs were similar in endosperm of diploids (2 X 2) and tetraploids (4 X 4). However, the 24-nt siRNA population was ~9% lower in 2 X 4 (37%) than in 4 X 2 (46%) seeds, but not in leaves (Figure 2.2D). The 24-nt siRNA densities in a 10-kp sliding window were significantly lower in 2 X 4 than in 4 X 2 seeds (Wilcoxon paired ranks sum test, $P = 0$), whereas the 24-nt siRNA density difference between 2 X 4 and 4 X 2 crosses in leaves was insignificant ($P = 0.2$).

Ploidy-dependent siRNAs are derived from transposable elements (TEs) and TE-associated genes (TAGs) in endosperm.

Average distributions of small RNAs were 16.9% in TEs, 19.5% in genes, 49.1% in intergenic regions (IGRs), and 14.4% in microRNAs (miRNAs) and trans-acting siRNA (ta-siRNAs) (Figure 2.3B). The 24-nt siRNAs were enriched in TEs and intergenic regions, whereas 21-nt siRNAs were derived from miRNA and tasiRNA loci. Consistent with the reduction of 24-nt siRNAs in 2X4 seeds, there was a lower proportion of small RNA reads in TEs (14%), genes (17%), and IGRs (45%) in 2 X 4 than in 4 X 2 seeds (19/21/54%).

By contrast, the fraction of miRNA and ta-siRNA reads was higher in 2 X 4 (24%) than in 4 X 2 seeds (7%). Among the up-regulated miRNAs and tasiRNAs in 2 X 4 seeds, many were from a few loci with abundant reads in seeds (Figure 2.4A-E). Correspondingly we demonstrated that miR832 but not miR172 and miR396 accumulated to higher levels in 2X4 seeds than in the other samples (Figure 2.3E). However there was no correlation between expression levels of miRNAs and tasiRNAs and their targets in 2

X 4 and 4 X 2 seeds (Figure 2.4E-I). The role for these miRNAs in seed development is not clear.

Interestingly, certain siRNA loci were markedly over- or underrepresented in these samples. Among 3,901 annotated TE genes, 24-nt siRNA densities in a 100-bp sliding window were 42% lower in 2X4 than in 4X2 seeds (Figure 2.3C) (Wilcoxon paired ranks sum test, $P = 0$). The 24-nt siRNA densities were 20% lower in tetraploids than in diploids ($P = 0$), suggesting a dosage compensation mechanism for siRNA expression in seeds that are larger in tetraploids than in diploids. Among 418 TE families each generating 100 or more unique 24-nt siRNA reads, most siRNAs were expressed at significantly lower levels in 2 X 4 than in 4 X 2 seeds ($P = 0$, Wilcoxon paired rank-sum test, Figure 2.3C). By contrast, 24-nt siRNA densities in leaves did not show a significant difference between 2 X 4 and 4 X 2 crosses (Figure 2.3D) ($P = 0.7$). The data suggest parent-of-origin effects of siRNAs on TE genes in developing seeds.

Corresponding to siRNA accumulation in TE genes, there was also an effect of siRNAs on TE-associated genes (*TAGs*). We mapped genomic coordinates of 31,189 TEs and TE fragments onto transcribed regions and 2-kb regions upstream and downstream of 27,379 protein-coding genes (TAIR9). We found that 12,676 protein-coding genes (~46%) contained at least one TE or TE fragment (Figure 2.5A). Among them, ~19% contained TEs in the 5' upstream (~2-kb), ~13% in the 3' downstream (~2-kb), ~12% in both upstream and downstream, and ~2% in the transcribed regions. Interestingly, siRNA densities in 5' and 3' regions of *TAGs* were significantly lower in 2 X 4 than in 4 X 2 seeds (Figure 2.5B, $P = 2.20E-16$, Wilcoxon paired rank-sum test), but not in leaves (Figure 2.5C, $P = 0.04$).

Does the reduction of siRNAs in 2 X 4 relative to 4 X 2 seeds affect gene expression in endosperm? To address this question we first used published microarray

data in reciprocal 2 X 4 and 4 X 2 or 2 X 6 and 6 X 2 crosses in young siliques (Tiwari et al., 2010), and we identified 9,742 *TAGs* and 13,029 non-*TAGs* on the array data. There were 151 *TAGs* that were differentially expressed between 4 X 2 and 2 X 4 siliques and, of those, 83% were more abundant in 2 X 4 than in 4 X 2 crosses. In contrast, of 90 non-*TAGs* only 67% were more abundant in the 2 X 4 cross (χ^2 = 19.13, P = 1.2 X 10^{-5}) (Figure 2.5D). The same trend was observed in reciprocal tetraploids (2 X 6 and 6 X 2) between diploid and hexaploid lines (χ^2 = 17.37, P = 3.1 X 10^{-5}) (Figure 2.5E). These data indicated a tendency of increased expression of *TAGs* in siliques of paternal-excess crosses. These *TAGs* could be expressed in endosperm together with the maternally expressed p4-siRNAs (Mosher et al., 2009), and their up-regulation in the paternal-excess crosses could be due to the reduced level of these siRNAs.

In our second approach to address the possibility that maternal p4-siRNAs are mediators of silencing in the endosperm we exploited previous analysis that had identified endosperm-preferred early seed stage (EP-ESS, also known as (aka) endosperm transcription factor, ETF) genes and silique tissue-preferred early seed stage (OST-ESS, aka silique transcription factor, STF) genes (Day et al., 2008; Le et al., 2010; Tiwari et al., 2010). The *ETF* and *STF* genes included 779 and 448 *TAGs*, respectively. Of the *ETF* genes, 60 were up-regulated, and 2 were repressed in 2 X 4 seeds relative to 4 X 2 seeds (Figures 2.6A and B). In contrast, only 4 *STF* genes were up-regulated and 1 was repressed in 2 X 4 seeds. Similar results were obtained from interploidy crosses between 2 X 6 and 6 X 2. Overall, the differentially expressed genes in interploidy crosses were enriched in gene ontology groups of hydrolase, receptor binding, and transcription factor activities (Figure 2.6C).

These data indicate that up-regulation of *ETF* genes is correlated positively with increased paternal genome dosage and negatively with increased maternal genome

dosage (hypergeometric test, $P = 0$). Moreover, among 27 *ETF* genes that are validated as regulators of endosperm development (Day et al., 2008), the majority were up-regulated in 2 X 4 seeds in two biological replicates (Table 2.3 and Figure 2.5F). No obvious trend was found among 25 *STF* genes (Figure 2.5G). Remarkably, 20 of 27 (~74%) *ETF* genes are *TAGs*, and they generated siRNAs (Table 2.3). In comparison, of all other genes, only ~46% are *TAGs* that generate siRNA ($P = 0.0001$). Up-regulation of these *EFT* genes may play a role in over-proliferation of endosperm in the paternal-excess triploids (Scott et al., 1998; Tiwari et al., 2010).

Maternal siRNAs regulate *AGAMOUS-LIKE* (*AGL*) gene expression.

The *ETF*s include a large family of *AGL*s that are members of the plant type I MADS domain subfamily (Parenicova et al., 2003). Many *AGLs* are expressed in female gametophyte or developing seeds and play a role in reproductive development (Bemer et al., 2010). Indeed, most *AGLs* were highly induced in seeds 3-8 days after pollination (Figure 2.7).

We next tested whether p4-siRNAs are associated with expression of *AGLs*. Thirteen of 61 (~21%) type I genes and five of 45 (~11%) type II genes, respectively, contained 10 or more normalized p4-siRNA reads in upstream, transcribed, and downstream regions (Table 2.4). All type I genes that generated siRNAs belong to Mα, Mγ and Mδ subgroups (Figure 2.8) that are expressed predominately in developing endosperm. The Mβ-type genes are mostly expressed in female gametophyte (Bemer et al., 2010), and none of them was found to be associated with p4-siRNAs. Type II genes such as *AGL42* are expressed in vegetative tissues and during floral transition (Dorca-Fornell et al., 2011).

Interestingly, the expression levels of all 13 siRNA-containing *AGLs* were higher in 2 X 4 than in 4 X 2 seeds (Figures 2.9A-G and 2.10), as observed in siliques (Tiwari et al., 2010). The transcript levels of these *AGLs* were inversely correlated with siRNA levels that were lower in 2 X 4 than in 4 X 2 seeds ($R^2 = 0.67$, $P = 0.025$) (Figure 2.6D).

We also tested the expression of additional *ETF* genes including six known imprinted genes and four candidate imprinted genes (Table 2.3), many of which overlapped with TEs that generated 24-nt siRNAs in 5' or 3' regions. Some without obvious TEs generated 24-nt siRNAs from their 5' upstream regions, suggesting presence of TE fragments or repeats that have not been annotated. The expression levels of these genes were higher in 2 X 4 than in 4 X 2 seeds but not obviously correlated with siRNA densities, with one exception. *FWA* is an imprinted gene in endosperm (Kinoshita et al., 2004). The maternally expressed *FWA* was up-regulated in paternal-excess seeds (2 X 4) (Figure 2.9H), which correlated with lower siRNA densities in endosperm in 2 X 4 than in 4 X 2.

Expression of p4-siRNAs and *AGLs* in interploidy crosses is dependent on PolIV.

Biogenesis of p4-siRNAs is, by definition, dependent on RNA polymerase IV (PolIV), a homologue of RNA polymerase II (Herr et al., 2005; Mosher et al., 2009; Onodera et al., 2005). *NRPD1a* encodes the largest subunit of PolIV. Absence of maternal *NRPD1a* reduces or eliminates 24-nt siRNA expression in endosperm (14). To test the effects of *NRPD1a* on *AGL* expression, we crossed a diploid *nrpd1a* (2*nrpd1a*) mutant (Mosher et al., 2009) with diploid (2x) and tetraploid (4x) wild-type plants in Col-0. *NRPD1a* expression was lower in 2*nrpd1a*X4 and 2*nrpd1a*X2 seeds than in corresponding reciprocal hybrids (Figure 2.11A) and a previously reported p4-siRNA,

siR02 (Mosher et al., 2009), was undetectable in 2*nrpd1a*X4 and 2*nrpdl1a*X2 seeds (Figure 2.11E).

Absence of maternal *NRPD1a* transcripts in these mutant crosses was associated with upregulation of 11 *AGLs* including *AGL40*, *AGL62*, and *AGL91* that were 2-3-fold higher in *2nrpd1a*x4 than in 2 X 4, as well as in 2*nrpd1a* X 2 seeds than in 2 X 2*nrpdl1a* (Figures 2.11*B-D* and 2.12A-C). Up-regulation of *AGL91* and *AGL40* in 2*nrpd1a* X 4 and 2*nrpdl1a* X 2 seeds correlated with down-regulation of p4-siRNAs (Figures 2.11E, right, and 2.12D). The p4-siRNAs associated with *AGL91* were present in seeds but not in siliques from which the seeds were removed (Figure 2.11E, left). A lower amount of 21-nt siRNAs was also associated with *AGL91*. The data indicate a link of maternal *nrpd1a* repression with reduction of maternal siRNAs and up-regulation of *AGLs*.

DISCUSSION

Our data collectively suggest a new model that explains the role for PolIV-dependent maternal siRNAs in *AGL* expression and endosperm development (Figure 2.11*F*). Proper seed development requires an endosperm balance number of 2m:1p in diploids (2 X 2) (Moore and Haig, 1991; Scott et al., 1998). In the maternal-excess endosperm (4 X 2, 4m:1p), maternal p4-siRNA expression levels increase, and p4-siRNA associated *AGLs* are repressed, causing precocious cellularization of endosperm and development of smaller seeds. In contrast, in the paternal-excess endosperm (2 X 4, 2m:2p) a low abundance of maternal and p4-siRNAs leads to up-regulation of *AGLs*, promoting endosperm nuclear proliferation and enlarging seeds.

Genome-wide demethylation in endosperm (Gehring et al., 2009; Hsieh et al., 2009) may lead to production of p4-siRNAs that are dependent on maternal genome dosage. During endosperm development, these maternal p4-siRNAs may directly silence

AGL targets and TEs, as observed in this study or indirectly through a mechanism of RNA-directed DNA methylation (Law and Jacobsen, 2010; Matzke et al., 2007). It is unclear whether these maternal siRNAs present in endosperm affect TEs and gene expression in embryo. There is a predicted movement of 21-nt siRNAs in vegetative and sperm nuclei of pollen (Slotkin et al., 2009). In addition, some of these siRNAs, most likely 21-nt siRNAs (Table 2.4), may participate in the post-transcriptional silencing pathway and trigger the secondary siRNA cascades through a tasiRNA-like mechanism (Chen et al., 2007; Montgomery et al., 2008). Indeed, *AGL91*, *AGL40* and *AGL36* had a significantly high probability ($p < 0.01$) of generating 21-nt phased siRNAs (Table 2.4) (Chen et al., 2007).

The endosperm and seed size is critical to the fitness of plants. Alteration in seed size is a manifestation of parental genome conflict in plants (Moore and Haig, 1991). Our model is consistent with that feature of parental genome conflict because the p4-siRNAs and their *AGL* transcription factor targets are all expressed in the endosperm. Endosperm is a triploid that contains two maternal (central cell) and one paternal (sperm) genomes. During the evolution of angiosperms, these maternal p4-siRNAs regulate expression of genes such as *AGLs* which are important to metabolism and nourishing function of maternal tissues (Moore and Haig, 1991; Stebbins, 1976). We also predict that the maternal p4-siRNAs serve as the factor for balancing and recognizing heterologous maternal and parental genomes in hybrids (Ng et al., 2012). The p4-siRNAs would allow the differentiation between paternal and maternal alleles, which could relieve the repressive maternal effects on the hybrids resulting from the parental genome imbalance and conflict in gametogenesis, fertilization, and early zygotic development (Bourc'his and Voinnet, 2010; Ng et al., 2012). The imbalance between the maternal siRNAs and their target genes in endosperm could lead to endosperm failure, a common syndrome

observed in many interspecific hybrids, but the embryos are viable and can be rescued to regenerate plants under tissue culture conditions (Sharma et al., 1996). These predicted effects could be readily tested in the hybrids within and between species.

MATERIALS AND METHODS

Plant materials and growth conditions

Diploids ($2n = 2x = 10$) and tetraploids ($2n = 4x = 20$) of *Arabidopsis thaliana* Col-0, C24, and L*er* ecotypes were grown under 16h light at 22°C and 8h darkness at 20°C. Reciprocal interploidy crosses were made by pollinating diploid flowers with tetraploid pollens (2 X 4) or tetraploid flowers with diploid pollens (4 X 2) 24h after manual emasculation. Diploid and tetraploid flowers were manually self-pollinated to serve as balanced dosage controls. Seeds were manually dissected from the siliques at 3, 4, 5, 6, 7 days after pollination (DAP) to eliminate maternal tissue contamination. Small RNA library construction and gene expression assays were performed using the seeds dissected at 6DAP. Rosette leaves of F_1 and their parents were collected before bolting for small RNA and gene expression studies.

Chromosome counts

A published protocol was adopted (Lysak et al., 2006). In brief, young floral buds were fixed in Carnoy's fixative (ethanol:glacial acetic acid, 3:1) and digested with pectolytic enzyme mixture (0.3% (w/v) cellulase, 0.3% (w/v) pectolyase and 0.3% (w/v) cytohelicase (all from Sigma) in citrate buffer) at 37 °C for 5 hours. Flower buds are then homogenized and spread on a glass slide by repeatedly adding 60% acetic acid and Carnoy's fixative. The chromosome spread was stained with 4',6-diamidino-2-phenylindole (Sigma) and examined under a widefield florescent microscope (Axiovert 200 M, Carl Zeiss). Three flower buds were examined per plant.

Small RNA library construction

Total RNA was extracted from seeds and leaves using Plant RNA reagent (Invitrogen) and subjected to electrophoresis in a 15% urea-polyacrylamide gel. The small RNA fraction (18-30-nt) was recovered from the gel. The small RNAs were ligated to 5' and 3' RNA oligo adapters (Table 2.5) and reverse-transcribed to produce first strand cDNAs, which were amplified by PCR and sequenced by Illumina Genome Analyzer II. Small RNA data are deposited in short read archives (http://www.ncbi.nlm.nih.gov/sra) (GSE25280).

Bioinformatic analysis

Short reads (40-nt) were parsed to remove 3' adaptors and mapped to *Arabidopsis thaliana* genome (TAIR9, June 2009 release) using CASHX (http://asrp.cgrb.oregonstate.edu/db/download.html) (Fahlgren et al., 2009). To reduce ambiguity, only the perfectly matched reads were used for further analysis. The sequences from chloroplast and mitochondrial and structural non-coding RNAs including ribosomal RNAs, transfer RNAs, snoRNAs and snRNAs were excluded from the analysis. Small RNA reads were normalized by library size and number of hits to the genome using the same weight for each matched locus.

Protein coding genes with adjacent TEs were identified by overlapping genomic coordinates of TEs and TE fragments with those of protein-coding genes using a Python script. Natural antisense gene pairs were defined as two loci overlapping with each other in the opposite direction. miRNA and tasiRNA targets were downloaded from ASRP database (http://asrp.cgrb.oregonstate.edu/db) (Gustafson et al., 2005). GOSlim terms were downloaded from http://www.arabidopsis.org/tools/bulk/go/index.jsp. The significance of enrichment was tested using hypergeometric test and Bonferroni multiple-testing correction (Holm, 1979) ($P = 0.01$).

Microarray analysis

The design and analysis of microarray datasets on interploidy crosses were described (Tiwari et al., 2010). Normalized data for 2 X 4, 4 X 2, 2 X 6 and 6 X 2 5DAP siliques (two replicates, Affymetrix) were downloaded from Gene Expression Omnibus (GSE20007). All statistic analyses were performed using R (http://www.r-project.org). Gene expression changes were estimated based on t-test between reciprocal crosses 2 X 4 vs. 4 X 2 and 2 X 6 vs. 6 X 2, respectively. Probe sets were called differentially expressed when $P \leq 0.01$ and log2-fold change ≥ 2.

Quantitative RT-PCR (qRT-PCR) analysis

Total RNA was extracted from leaves and dissected seeds using Plant RNA reagent (Invitrogen) and treated with DNaseI (Promega, Madison, WI). First-strand cDNA was synthesized using SuperScriptIII Reverse Transcriptase (Invitrogen, Carlsbad, CA). Primer sequences are listed in Table 2.6. Actin 2 (*ACT2*) was used as the internal control. qRT-PCR is performed using Applied Biosystems 7500 Real-Time PCR Systems (ABI, Foster City, CA).

Small RNA blot analysis

Small RNA (<200nt) was enriched from 50 mg total RNA using mirVana miRNA purification kit (Ambion). Small RNA was separated on a 15% denaturing 19:1 acrylamide:bisacrylamide gel with 1XTBE and 7M urea and transferred to HybondN$^+$ membrane (GE/Amersham). Oligonucleotides were end-labelled with [γ-^{32}P]ATP and T4 polynucleotide kinase (NEB) and hybridized with the membrane in Church's buffer (Church and Gilbert, 1984) at 37°C overnight. A mixture of oligonucleotides corresponding to the most abundant siRNAs from each *AGL* locus in the sequencing libraries was used as the probe to detect *AGL*-related siRNAs (Table 2.7). The blots were

washed twice in 2xSSC, 0.1% SDS at 50°C before exposure to phosphor-storage screens. Probe sequences are listed in Table 2.7.

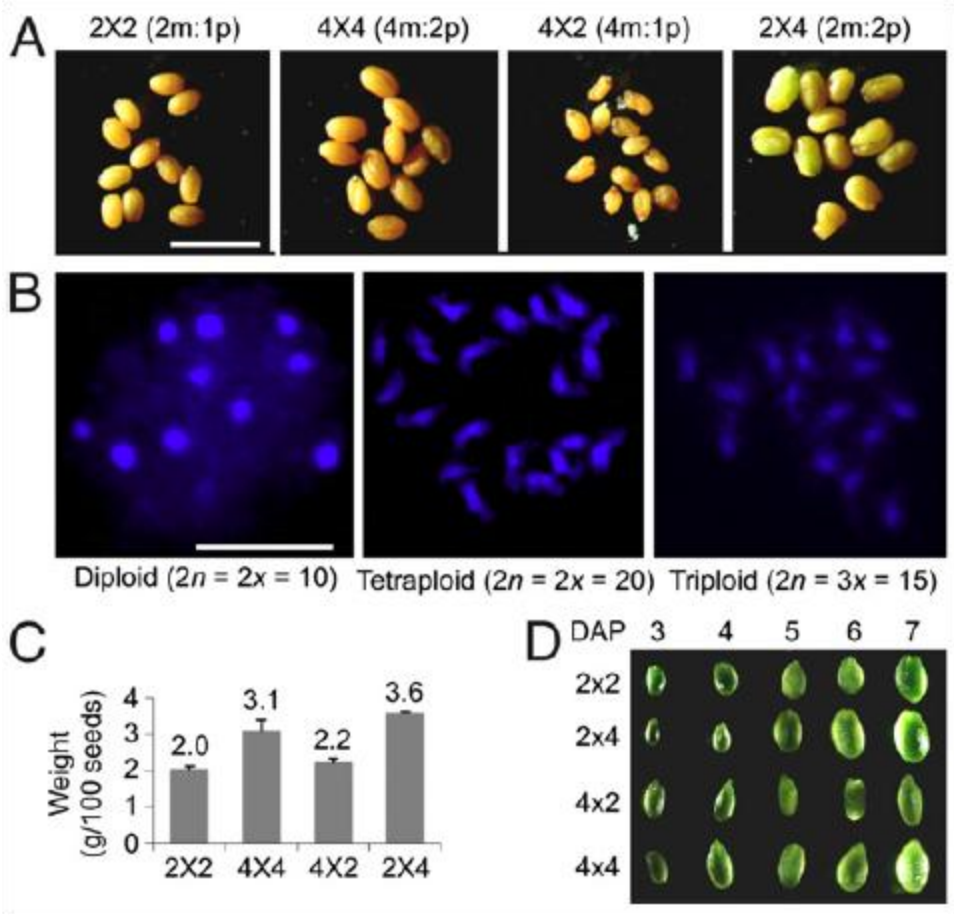

Figure 2.1: Seed morphology and chromosome counts in interploidy crosses. (A) Seed size and morphology in diploids, tetraploids and triploids (2 X 4 or 4 X 2) in *A. thaliana* C24. m: maternal; p: paternal. By convention, the maternal parent is listed first in a genetic cross. (B) Chromosome counts in diploid, triploid, and tetraploid flowers in *A. thaliana* Col-0. (C). Seed weight in interploidy crosses. Error bars were derived from three biological replicates. (D) Developing seeds dissected at 3-7 days after pollination in Col-0 (DAP).

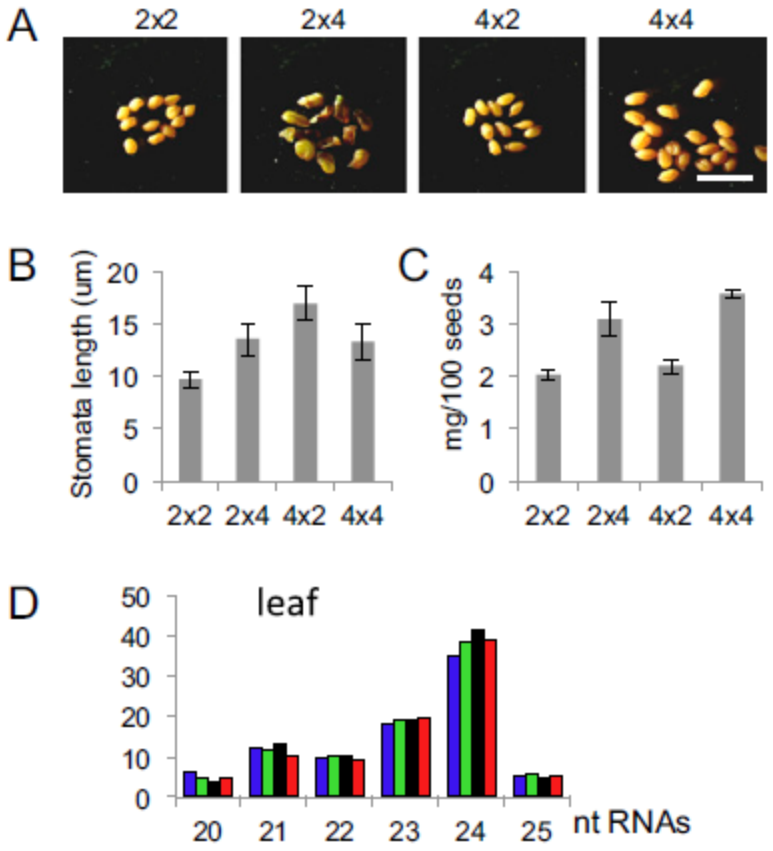

Figure 2.2: Seed size variation in ploidy series (Col-0) and small RNA distribution in leaves. (A) Seed size variation in ploidy series of Col. Note that seed abortion in late stage of development in Col is related to maternal expression of *TTG*2 (Dilkes et al., 2008, PLoS Biology 6:2707-2720) and possibly other unknown genes. Scale bar indicates 1mm. (B) Quantification of stomata sizes in ploidy series. (C) Quantification of dry seed weight in ploidy series. (D) Distribution of small RNAs in leaves.

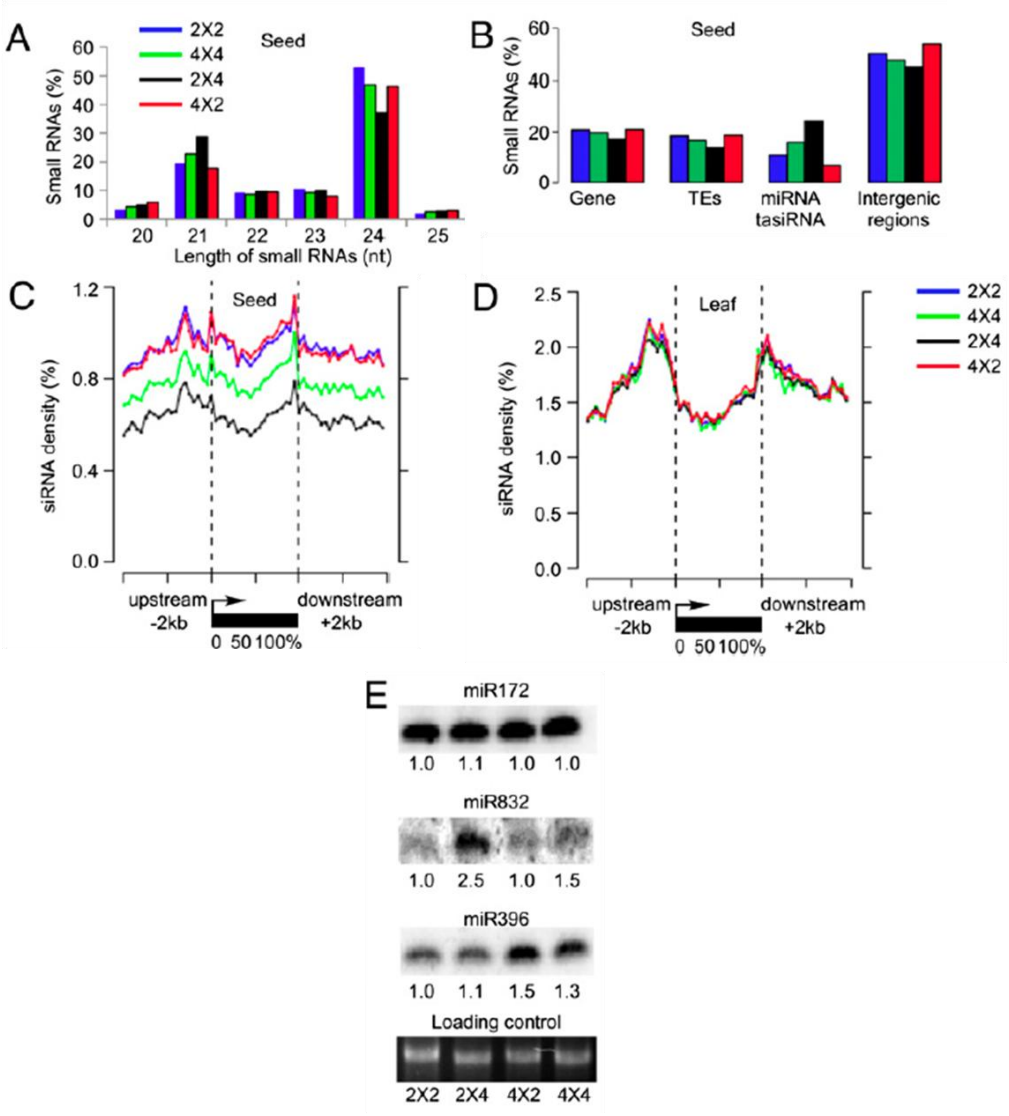

Figure 2.3: Small RNA distribution in interploidy crosses. (A) Size distribution of 20-25-nt small RNA reads in Col-0 seeds of 2X2 (blue), 4X4 (green), 2X4 (cyan) and 4X2 (magenta). (B) Distribution of 20-25-nt small RNAs in genes, TEs, miRNA and tasiRNA targets, and intergenic regions. (C-D) 24-nt small RNA densities (100-bp sliding window) in 5' upstream (2-kb), transcribed, and 3' downstream (2kb) regions of TE genes in seeds (C) and leaves (D). (E) Small RNA blot analysis of miR172, miR832, and miR396 in diploids, triploids, and tetrpaloids at 6 DAP.

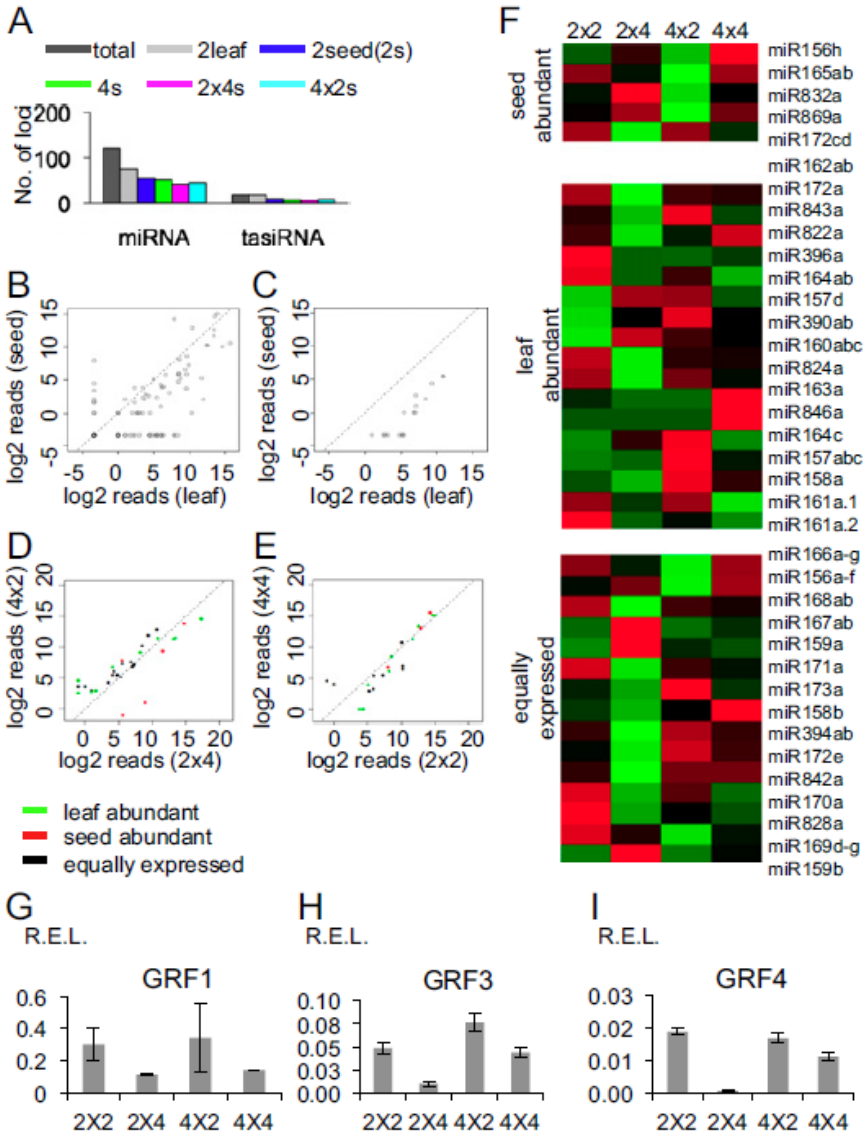

Figure 2.4: miRNA and tasiRNA expression in ploidy series. (A) Number of miRNA and tasiRNA loci identified in leaves and seeds. (B) miRNA expression in seed and leaf. (C) tasiRNA expression in seed and leaf. (D) miRNA expression in interploidy crosses. (E) miRNA expression in balanced crosses. (F) Heatmap of differentially expressed miRNAs in ploidy series seeds. (G-I) Association of TEs with siRNA abundance in protein-coding genes, *GRF1* (AT4G09000) (G), *GRF3* (AT5G38480) (H), and *GRF4* (AT3G52910) (I).

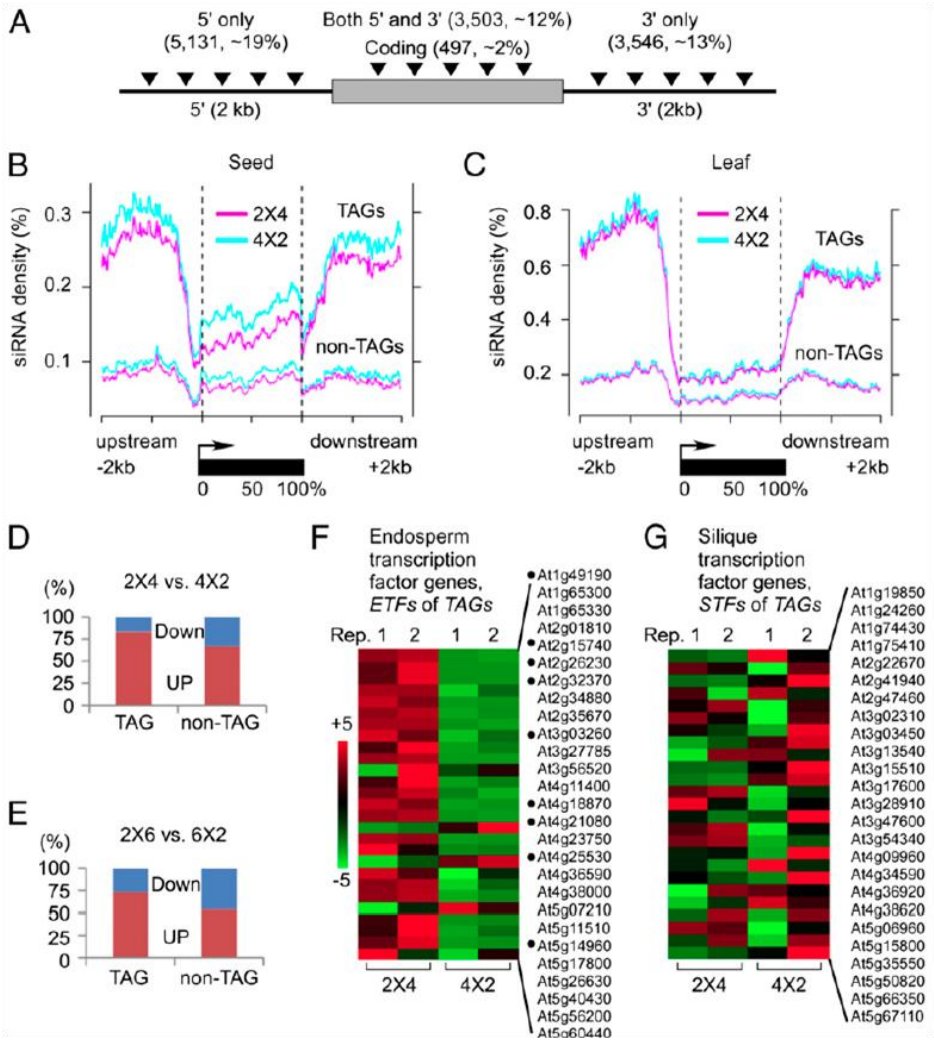

Figure 2.5: Distribution of TE-associated genes (*TAG*s) and parent-of-origin effects of siRNAs on *TAG* expression in reciprocal interploidy crosses. (A) Proportions of *TAG*s and locations of TEs (triangles) in coding sequences (grey box) and within 2-kb up or downstream of 5' and 3' regions (extended lines) in *A. thaliana* Col-0. (B-C) Small RNA densities (100-bp sliding window) in 5' upstream (2-kb), transcribed, and 3' downstream (2kb) regions of *TAG*s and non-*TAG*s in seeds (B) and leaves (C). (D-E) Percentage of upregulated (red) and downregulated (blue) *TAG*s or non-*TAG*s in 2 X 4 vs. 4 X 2 crosses (D) and in 2 X 6 vs. 6 X 2 crosses (E). (F-G) Heatmaps of gene expression changes in endosperm transcription factor (*EFT*) genes (n = 27) (F) and silique transcription factor (*STF*) genes (n = 25) (G) in two replicated experiments (Rep. 1 and 2); color bar indicates up (red) and down (green) regulation; black dots indicate upregulated genes at statistically significant levels ($P < 0.05$) between 2 X 4 and 4 X 2 endosperm.

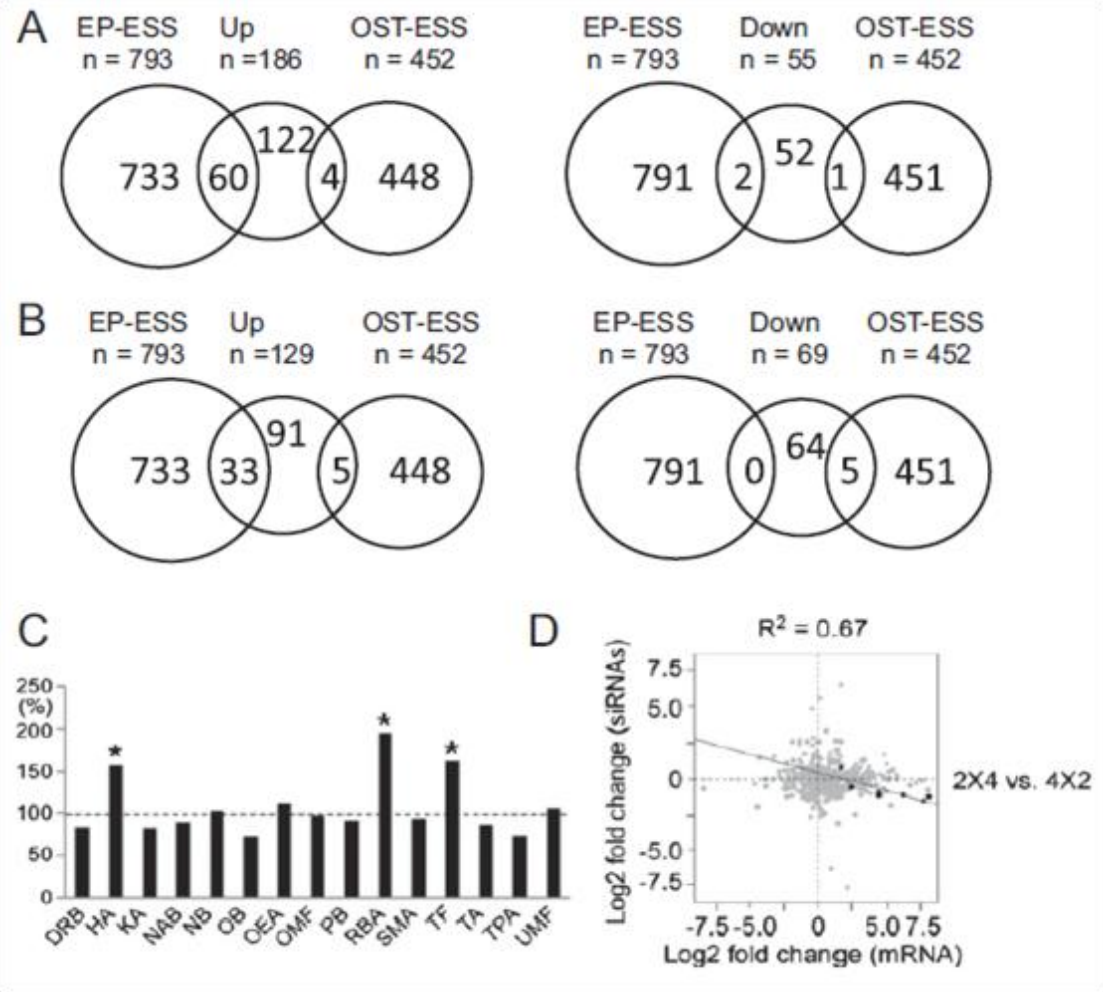

Figure 2.6: Correlation of differentially expressed genes and siRNAs in interploidy crosses. (A-B) Venndiagram of EP-ESS and OST-ESS with upregulated genes (left) and downregulated genes (right) in 2 X 4 seeds relative to 4 X 2 seeds (A) or in 2 X 6 seeds relative to 6 X 2 seeds (B). (C) GOSlim term showing an enrichment of siRNA generating genes; stars indicate GO groups with $P < 0.01$; DRB: DNA or RNA binding; HA: hydrolase activity; KA: kinase activity; NAB: nucleic acid binding; NB: nucleotide binding; OB: other binding; OEA: other enzyme activity; OMF: other molecular functions; PB: protein binding; RBA: receptor binding or activity; SMA: structural molecular activity; TF: transcription factor activity; TA: transferase activity; TPA: transporter activity; UMF: unknown molecular functions. (D) Inverse correlation of mRNA and siRNA abundance with expression levels of *AGLs* (black) but not with that of other siRNA generating genes (grey) in reciprocal interploidy crosses (2 X 4 vs. 4 X 2).

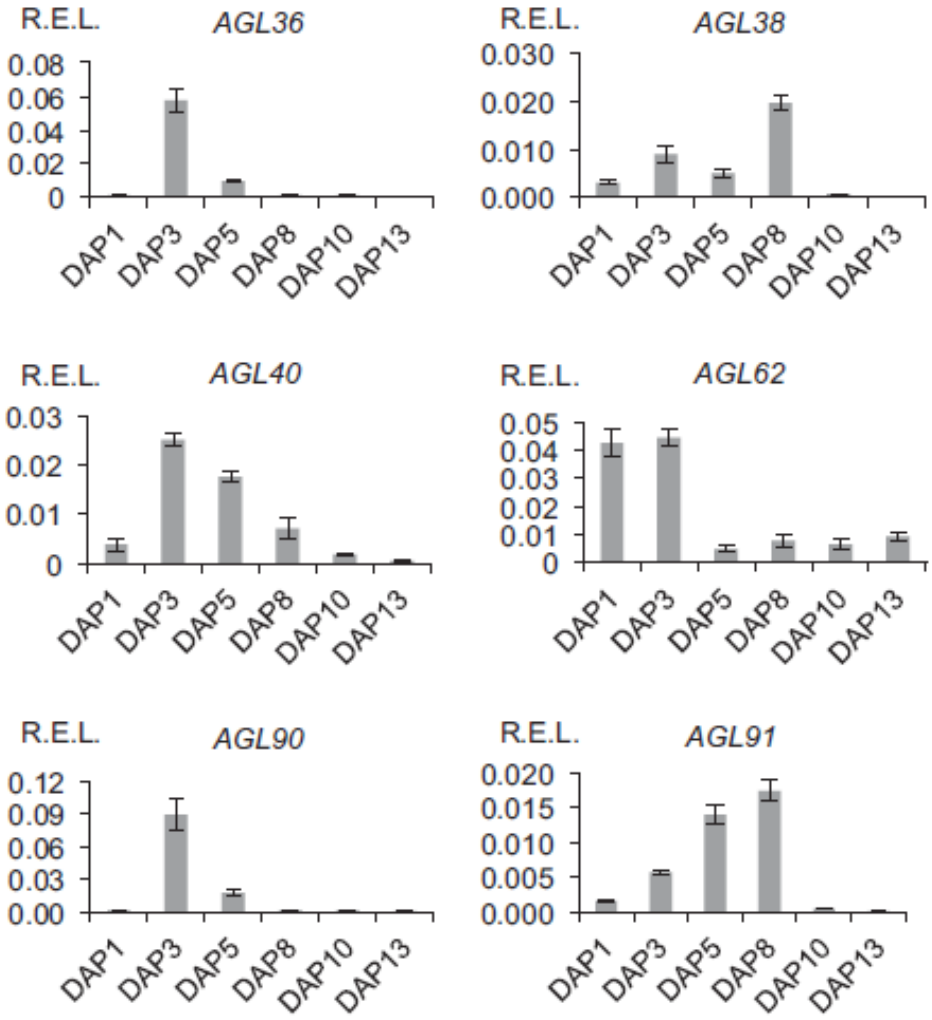

Figure 2.7: qRT-PCR analysis (REL, relative expression levels) of siRNA-related *AGL* genes in different stages of developing siliques.

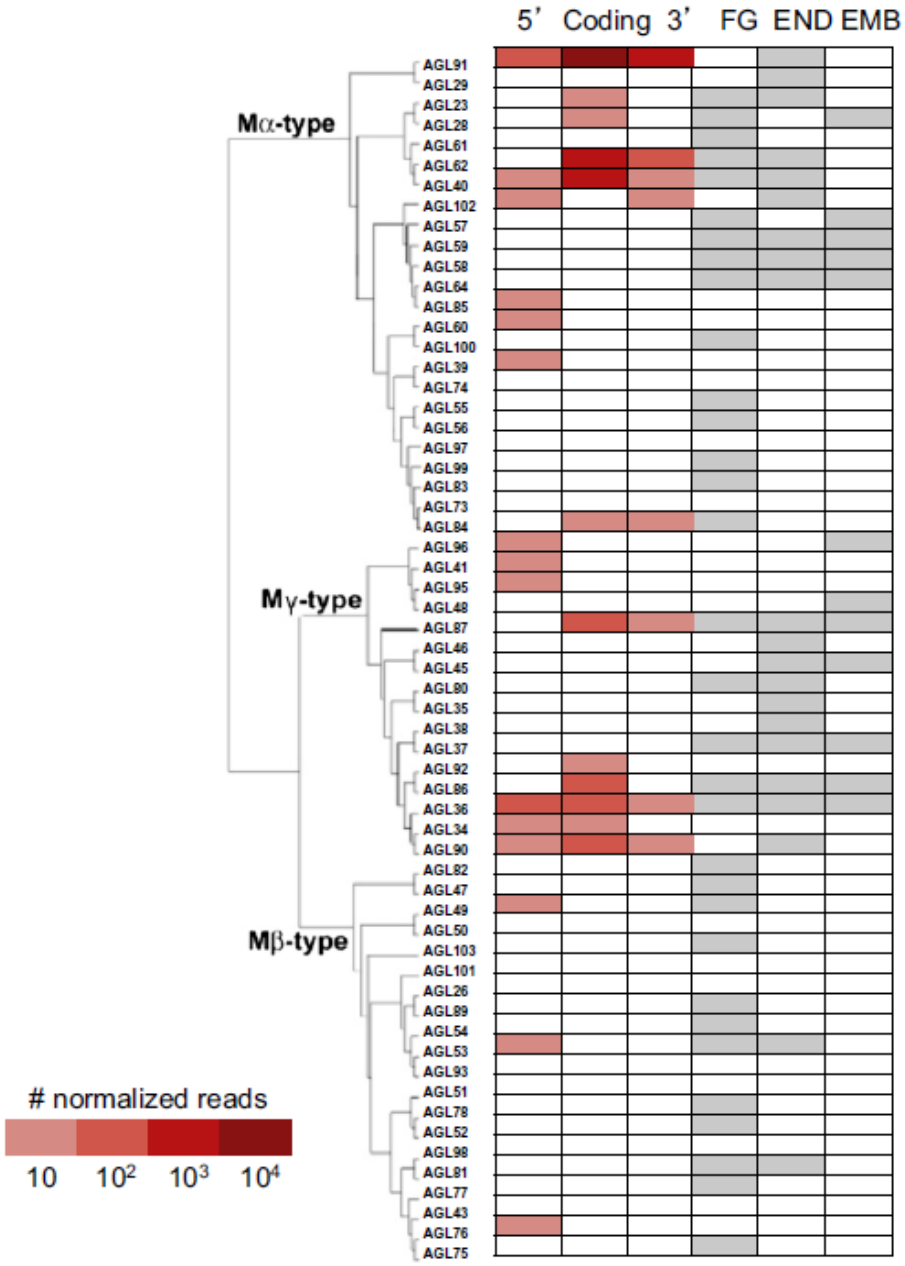

Figure 2.8: Small RNA expression in the type I MADS box transcription factors and the gene expression patterns during female gametophyte and seed development. The genes are arranged according to the phylogenic tree. Gradient colors represent different levels of small RNA expression measured by number of normalized 24-nt reads in this study. Up: upstream 2kb region of a gene; Gene: gene loci; Dn: downstream 2kb region of a gene. Spotted boxes indicate expression detected in other studies (Bemer et al., 2010; Day et al., 2008; Portereiko et al., 2006; Walia et al., 2009). FG: female gametophyte; END: endosperm; EMB: embryo.

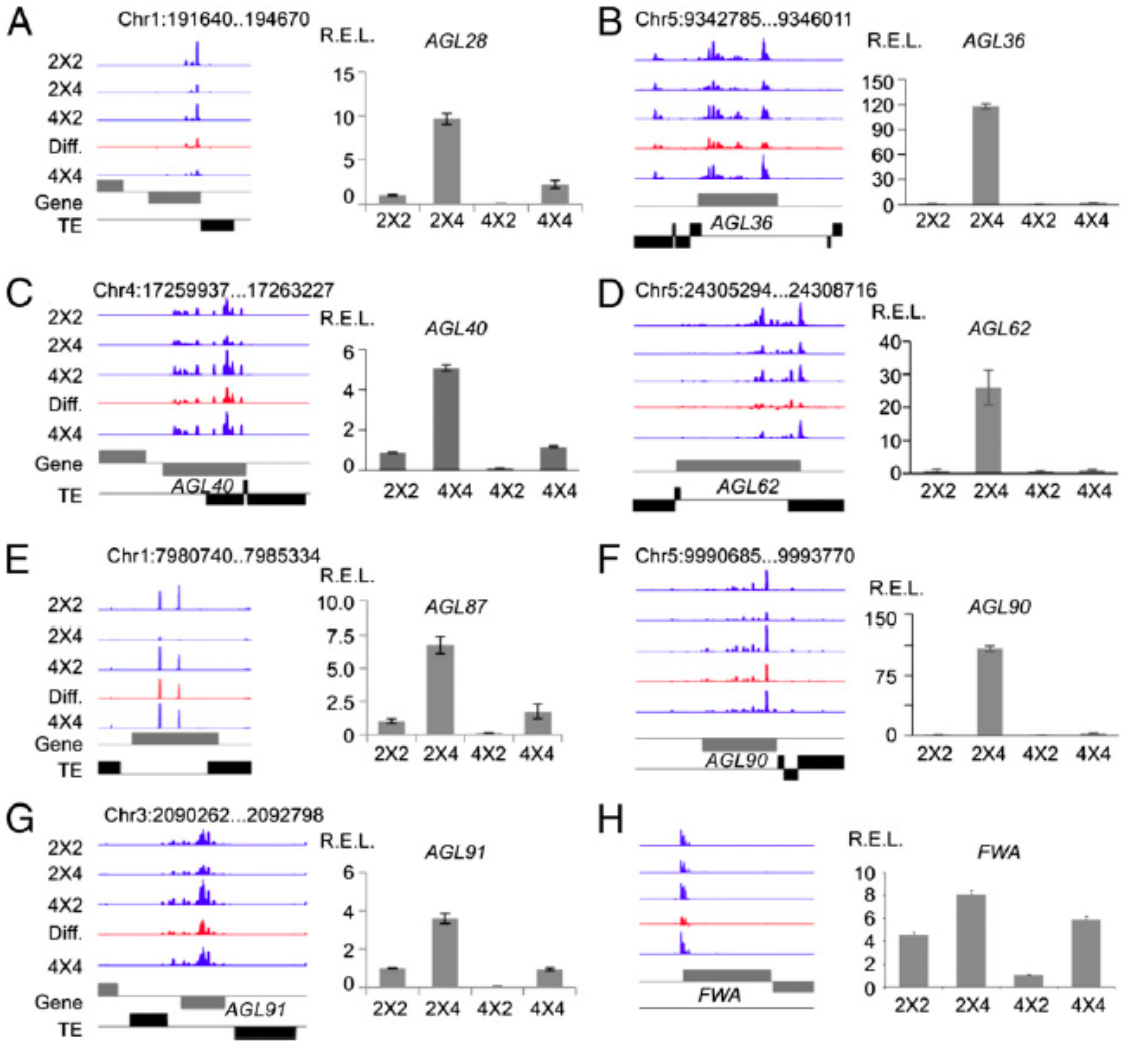

Figure 2.9: Maternal siRNAs are associated with expression of AGL genes and FWA. (A-H) siRNA hotspots (left) and qRT-PCR analysis (right) of *AGL28* (*A*), *AGL36 (B)*, *AGL40* (C), *AGL62* (D), *AGL87* (E), *AGL90* (F), *AGL91* (G), and *FWA* (H) in 2 X 4 and 4 X 2 triploids and their parents (2 X 2 and 4 X 4). Diff.: siRNA differences between 4 X 2 and 2 X 4; positive: above the line; negative: below the line; grey box: gene; black box: transposon. Genomic coordinates are shown above each diagram, and standard errors were calculated from three biological replicates.

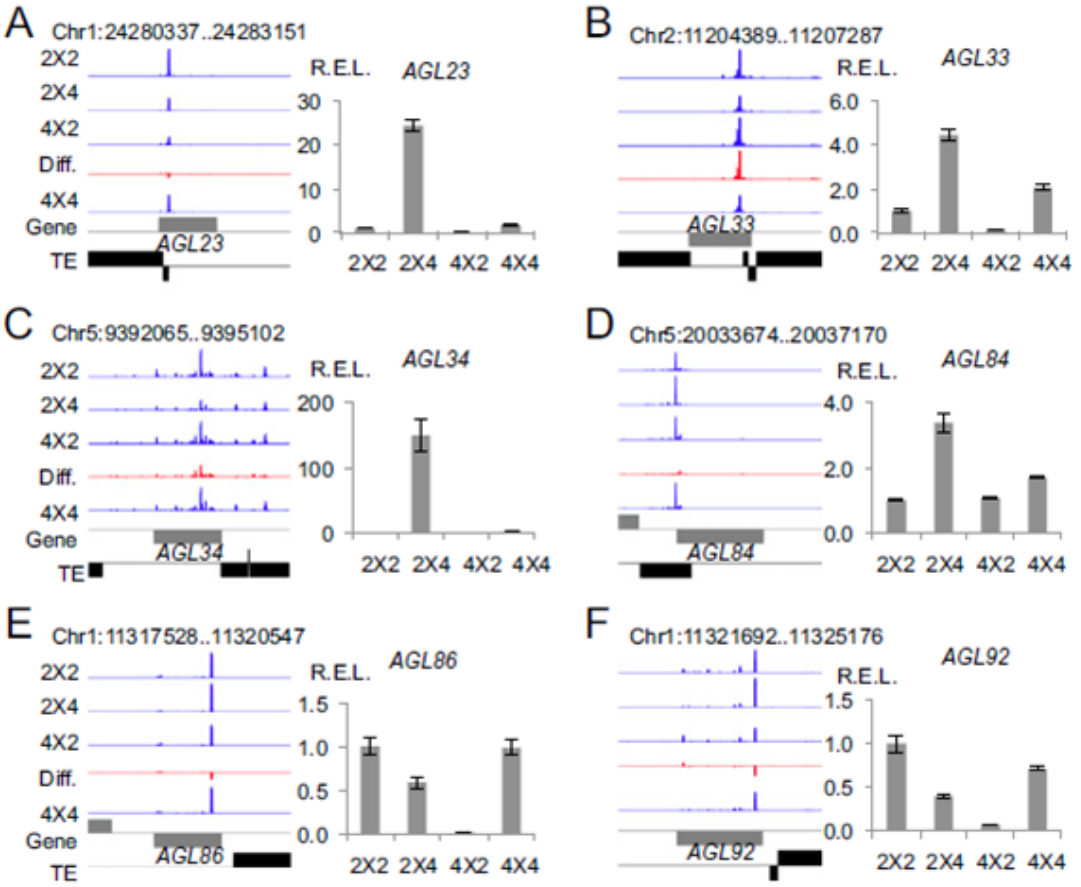

Figure 2.10: siRNA profiles (Left) and qRT-PCR analysis (Right) of additional *AGL*s in 2 X 4 and 4 X 2 triploids and their parents (2 X 2 and 4 X 4). (A) *AGL*23, (B) *AGL*33, (C) *AGL*34, (D) *AGL*84, (E) *AGL*86, and (F) *AGL92*. Diff, siRNA read differences that were higher in 4 X 2 than 2 X 4 (above the line) or vice versa (below the line); grey box, gene; black box, transposon. Genomic coordinates are shown *Above* each diagram. SEs were calculated from three biological replicates.

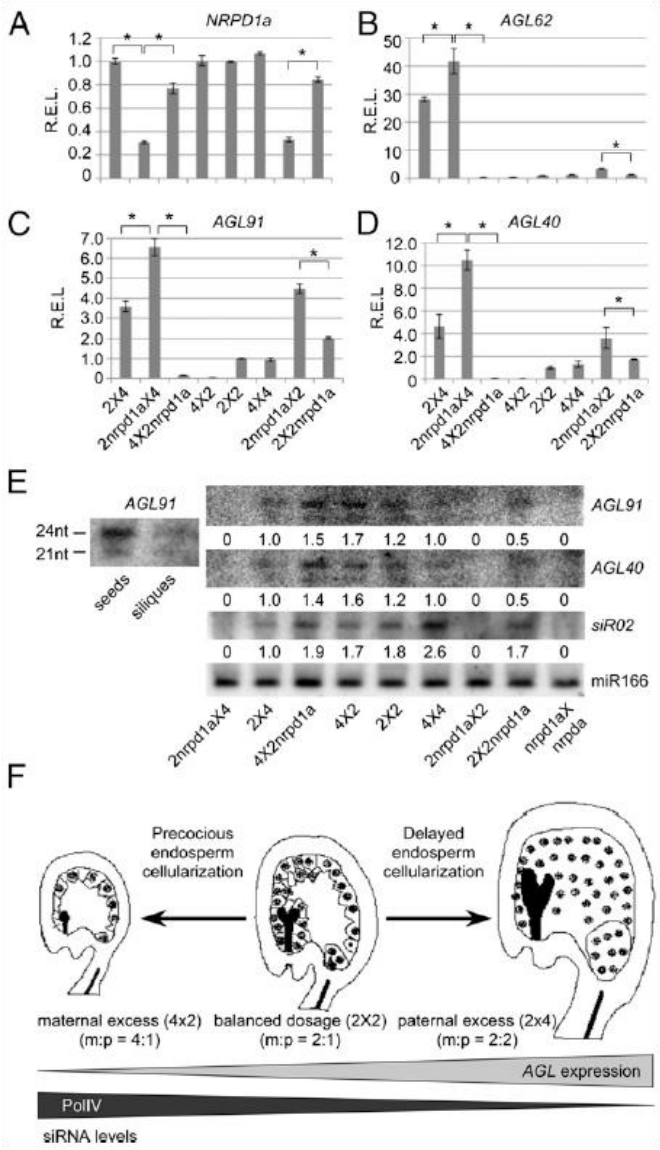

Figure 2.11: siRNA production and *AGL* expression are dependent on RNA polymerase IV (NRPD1A) in endosperm and a model for endosperm development in interploidy crosses. (A) qRT-PCR analysis (relative expression levels, R.E.L.) of *NRPD1a* expression in endosperm of interploidy crosses. (B-D) qRT-PCR analyses of *AGL62* (B), *AGL91* (C) and *AGl40* (D) expression (n = 3 biological replicates). (E) Small RNA blot analysis of p4-siRNA (siR02), *AGL40*-siRNA, and *AGL91*-siRNA in endosperm of interploidy crosses (n = 2 biological replicates). Left: *AGL91*-siRNAs were present in seeds but not in siliques. miR166 was used as a control. (F) A model for the role of maternal siRNA-mediated *AGL* expression in endosperm and seed development (see text for explanation). Multiple dots and an elongated black rod in each diagram represent the endosperm and embryo cells, respectively.

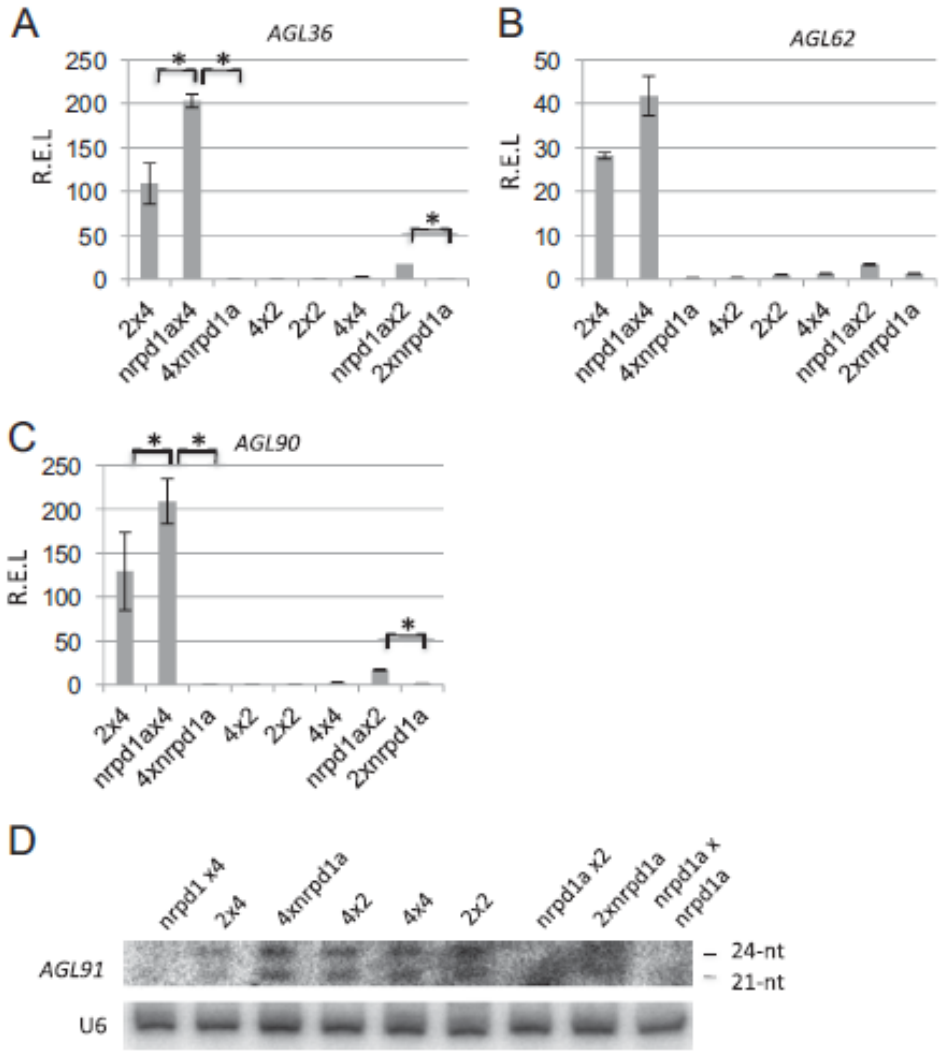

Figure 2.12: qRT-PCR (REL, relative expression levels) and small RNA blot analysis in interploidy crosses with *nrpd1a*. (A-C) R.E.L. of *AGL36* (A), *AGL62* (B) and *AGL90* (C) (n = 3 biological replicates). (D) Small RNA blot analysis of *AGL*-siRNAs. Both 24- and 21-nt siRNAs were detected in *AGL91*. U6 was used as a control.

Table 2.1: Sequence reads of small RNA libraries from the leaves of *Arabidopsis thaliana* reciprocal interploidy crosses and their parents[2]

Genomic feature	2L(%)	4L (%)	2X4L(%)	4X2L(%)
all reads	8,929,273	3,165,346	11,610,370	11,886,105
filtered	4,257,442 (48)	895,774 (28)	3,072,403 (26)	4,697,603 (40)
gene	896,187 (10)	360,142 (11)	1,414,223 (12)	1,153,949 (10)
transposon	1,297,503 (15)	543,352 (17)	2,606,896 (22)	2,410,799 (20)
miRNA	890,159 (10)	829,203 (26)	2,479,074 (21)	1,030,473 (9)
tasiRNA	16,333 (0.2)	7,075 (0.2)	46,154 (0.4)	22,157 (0.2)
Intergenic regions	1,571,649 (18)	529,800 (17)	1,991,620 (17)	2,571,124 (22)
total small RNA	4,671,831 (52)	2,269,572 (72)	8,537,967 (74)	7,188,502 (60)

Table 2.2: Sequence reads of small RNA libraries from the seeds of *Arabidopsis thaliana* reciprocal interploidy crosses and their parents

Genomic feature	2S (%)	4S (%)	2X4S (%)	4X2S (%)
all reads	12,082,171	11,642,856	12,705,829	11,516,592
filtered	778,465 (6)	1,274,150 (11)	2,637,149 (20)	1,742,706 (15)
gene	2,330,912 (19)	2,039,493 (18)	1,729,226 (13)	2,029,347 (18)
transposon	2,069,358 (17)	1,740,812 (15)	1,398,336 (11)	1,814,304 (16)
miRNA	1,227,444 (10)	1,647,975 (14)	2,421,478 (19)	668,594 (6)
tasiRNA	469 (0.004)	550 (0.004)	395 (0.003)	688 (0.006)
intergenic regions	5,675,523 (47)	4,939,876 (42)	4,519,245 (36)	5,260,953 (46)
total small RNA	11,303,706 (94)	10,368,706(89)	10,068,680 (79)	9,773,886 (85)

[2] All reads: reads perfectly match the genome; filtered: number of reads that match chloroplast, mitochondrial genome and structural non-coding RNAs, including rRNA, tRNA, snoRNA and snRNA.

Table 2.3: The expression changes (log2 fold change) of 27 validated genes that encode endosperm preferred early seed stage transcription factors[3]

AGI locus	Name	Log2 FC	p value	Imprinted	Adjacent TEs
AT2G32370	*HDG3*	4.7	0.0001	P (Gehring et al., 2009)	ATREP10D
AT3G03260	*HDG8*	7.7	0.0010	M (Gehring et al., 2009)	HELITROY3,ATREP11
AT4G21080	*Dof zinc finger protein*	3.7	0.0012		HELITRONY3, HELITRONY1B,HELITRONY1D (X2)
AT4G25530	*FWA/HDG6*	3.6	0.0015	M (Kinoshita et al., 2004)	NA
AT1G49190	*ARR19*	3.2	0.0027		ATREP6, ATLINE1_3A
AT2G26320	*AGL33*	4.8	0.0030		HELITRONY3, HELITRONY1E(X2), ATREP10A,ATREP4, ATREP7
AT5G14960	*DEL2*	2.6	0.0043		NA
AT4G18870	*Heat shock transcription factor family*	4.4	0.0067		ATREP3
AT2G15740	*C2H2-type zinc finger protein family*	4.4	0.0073		HELITRONY1B(X2), HELITRONY1D, ATREP5, ATREP10D,BOMZH1
AT4G11400	*ARID/BRIGHT DNA-binding domain protein family*	3.4	0.0116	C (Gehring et al., 2009)	ATDNATA1
AT2G01810	*PHD finger protein family*	1.9	0.0164	C (Gehring et al., 2009)	ATCOPIA75, TSCL, ATHATN4
AT5G11510	*MYB3R4*	0.9	0.0242		NA

[3] Log2 FC: log2 fold change of gene expression (2 X 4 vs. 4 X 2); p value: the p value of t-test; imprinted: the gene is known as maternally imprinted (M), paternally imprinted (P) or is one of the candidate imprinted genes as predicted in Gehring, M., Bubb, K.L., and Henikoff, S. (2009). Extensive demethylation of repetitive elements during seed development underlies gene imprinting. Science 324, 1447-1451.; Adjacent TEs: transposable element genes or fragments within the ±2-kb regions of the gene.

Table 2.3, cont.

AT5G56200	*C2H2-type zinc finger protein family*	3.0	0.0246		ATCOPIA24, ATREP10D, BRODYAGA1A
AT1G65300	*PHE2/AGL38*	7.2	0.0277		HELITRONY1D, BRODYAGA1A
AT1G65330	*PHE1/AGL37*	7.2	0.0277	P (Kohler et al., 2005)	ATREP11
AT2G34880	*MEE27*	0.8	0.0289		ATCOPIA57
AT4G38000	*Dof zinc finger protein*	-0.7	0.0448		NA
AT2G35670	*FIS2*	2.0	0.0542	M (Luo et al., 2000)	NA
AT3G56520	*No apical meristem protein family*	3.6	0.0598		ATREP15, ATREP1
AT5G26630	*AGL35*	4.1	0.0602		ATREP10B, ATREP3, ATREP4, ATREP10D
AT5G40430	*MYB22*	1.4	0.0805	C (Gehring et al., 2009)	ATREP10D, ATDNA12T3_2
AT4G23750	*CRF2*	-1.5	0.0855		NA
AT5G07210	*ARR21*	1.3	0.1151	C (Gehring et al., 2009)	ATREP4 (X2), ATREP5, ATREP9, ATREP11, HELITRONY1E, BRODYAGA2
AT4G36590	*AGL40*	5.4	0.1170		ATREP10D, ATREP19, BRODYAGA2, BRODYAGA1A
AT5G17800	*MYB56*	-1.2	0.1360		NA
AT5G60440	*AGL62*	0.8	0.4808		ATREP4, ATREP15 (X2), HELITRONY3, BRODYAGA1
AT3G27785	*MYB118*	0.1	0.8740	C (Gehring et al., 2009)	ATREP4

Table 2.4: A list of MADS box genes that possess more than 10 reads from transcribed regions in at least one line[4]

Locus	Name	2X2	4X2	2X4	4X4	p-value	TE	Type
AGL91	AT3G66656	15029	17843	10688	22982	0.0079	3	I
AGL40	AT4G36590	7941	9395	4834	11665	0.0092	4	I
AGL33	AT2G26320	580	797	452	1076	0.0540	6	I
AGL36	AT5G26650	980	886	510	1034	0.0009	6	I
AGL86	AT1G31630	1000	950	1052	835	0.0200	2	I
AGL62	AT5G60440	1121	530	548	820	0.0393	5	I
AGL90	AT5G27960	252	252	122	311	0.0495	5	I
AGL34	AT5G26580	82	75	39	98	0.0849	5	I
AGL87	AT1G22590	117	100	5	84	0.2453	5	I
AGL92	AT1G31640	72	43	73	61	0.0712	5	I
AGL23	AT1G65360	84	55	40	54	0.2088	3	I
AGL28	AT1G01530	28	12	12	23	0.0893	3	I
AGL84	AT5G49420	11	9	7	15	0.0305	1	I
AGL42	AT5G62165	22	22	42	34	0.1165	0	II
AGL66	AT1G77980	32	12	10	13	0.0456	3	II
AGL14	AT4G11880	7	29	4	12	0.3756	1	II
AGL12	AT1G71692	20	11	24	10	0.2452	0	II
AGL11	AT4G09960	10	16	13	6	0.3756	4	II

Table 2.5: 5’ and 3’ adaptor sequences for small RNA library construction.

adaptor	sample	sequence
5' adaptor	2X2 seed/leaf	5'-GUUCAGAGUUCUACAGUCCGACGAUCA-3'
	4X4 seed/leaf	5'-GUUCAGAGUUCUACAGUCCGACGAUCT-3'
	2X4 seed/leaf	5'-GUUCAGAGUUCUACAGUCCGACGAUCC-3'
	4x2 seed/leaf	5'-GUUCAGAGUUCUACAGUCCGACGAUCG-3'
3' adaptor		5' P-UCGUAUGCCGUCUUCUGCUUGUidT

[4] The 24-nt reads were normalized per 10 million; %, percentage of reads from the sense strand; P values, probability of a locus that generates secondary siRNAs by chance; TE, number of transposons in the locus and ±2-kb regions.

Table 2.6: List of qPCR primers used for this study.

Gene	Forward Primer Sequence	Reverse Primer Sequence
AGL23	TGACCACTTTCGAGGGTGTGTTG	TTACTCCACCACTCCTCAGCGTTT
AGL28	ACTTGACCACTCTTTAGGGCGTG	ACCACTTCTCAGCGTCCTTGTTCT
AGL33	TTGTTTCTCCCACCGAGAAGCCT	TCACACCTCTCTTCCCGTTCTTGT
AGL34	GATGCGAATGCAACTGCGGTAAC	ACAAGGTGTTGAAACGGTCGATGC
AGL36	ATGAATCATGTTGGAGGGCG	GCAGTTTCCGTCCACGAAAG
AGL40	GCCATTGTCCATCATCAGAACA	CGGACGGGTTTCAACAAGTT
AGL42	GATCGAACGCTACCGCAAGT	TCGTGATTGCTGGTTTCATGA
AGL62	AATTGGTGGGAAGATCCCGT	CCCCTCGAGTTGAGATAACGC
AGL84	ACCAAACGCCGTGAAGGTCTCTA	AGACAACGGCATCAACAGAGGAG
AGL86	TGAGCATCAGTATGGTGCAAGGG	TGAAGCGGAGGAGGAAAGGAAGA
AGL87	ACCGTTCTATGTGGTTTACCCGCT	TGCTTCTCTACCGGAAGTTCGCTT
AGL90	ACCAGCCGTTGATCTTGCTT	ACATCGGTGGTTGAAGGCAT
AGL91	AACAAACCGTATTCCTTCGGG	CCGTTCTGCAATCACATCAAA
AGL92	TCATGAGACTTTCTTGCGGGACC	TGAAGGTCCCTTGCACCATACTGT
FIS2	TCTTGCCCATTTTGCTTGATT	AAGTTGCAAGCCCTCGTGAC
FWA	AGCCTGGTGAGCTAACTGGG	GCCAAACAGAAGTGGATGCAC
MEA	GTTTGGATGATCTGGTCGTGC	CCACTTCGAGGTACTTGGCG
PHE1	GTGGTGTTGACGCATGTGC	CCTGGATCGAGTTGTACGGG
GRF1	GGATTAGGCGTCAACACCGA	GTTATTCGTCTTTTTCCCGGG
GRF3	TGAGGCCCTTCTTTGACGAT	TGTCAGCTTCTTGGAGCGAA
GRF4	CACCAACCTTCTTGGTATTGGG	CCCTGGCTCAGGATCCATT
NRPD1A	GGCGGGTGAGCTGTACTTGA	CTTTTGCCCCGATCTCCATA

Table 2.7: Lists of probes used for this study.

Gene	Sequence
miR166	GGG GAA TGA AGC CTG GTC CGA
miR172	ATG CAG CAT CAT CAA GAT TCT
miR396	AGTTCAAGAAAGCTGTGGAA
miR832	TTTCGATTCCCGATCCCAGCA
AGL91	TTGCCTCTACTATAGCCTGAT; ATTGCCTCTACTATAGCCTGA; TCTTAAACCGTTCTGCAATCACAT; GTTGAGGCGTTTACATATCTTCTT; CTTAAACCGTTCTGCAATCACATC; CTATTGCCTCTACTATAGCCTGAT; TTGCCTCTACTATAGCCTGATGTT
AGL40	ATACCAGTTTCCTACTTGTTCTCT; ATGCCAATTCTATGTCTATAACTT; TTCTGCCATTGTCCATCATCA; AACCATAATTCTGCCATTGTCCAT; TTCATCGCTTTTCGAATCCTA; TGACGACCTTTGGTACTTCT
U6	GCTAATCTTCTCTGTATCGTTCC

Chapter 3: Maternal siRNAs Guide Spatiotemporal Regulation of Euchromatic Loci and DNA Methylation in *Arabidopsis* Endosperm[5]

ABSTRACT

In plants and animals, small interfering RNAs (siRNAs) mediate epigenetic inheritance of heterochromatin and genome stability. We found unique roles for maternal siRNAs in RNA-directed DNA methylation (RdDM) and expression of euchromatic loci in endosperm of seed. The majority of maternal siRNAs are derived from euchromatic regions containing short transposable elements. These euchromatic loci including many *AGAMOUS-LIKE* genes (*AGLs*) are subjected to RdDM in spatiotemporal-specific manners. Both *AGL91* and *AGL40* are actively expressed in chalazal endosperm where RdDM remains inactive up to the heart stage but silenced in other regions where RdDM is active. *AGL91* is paternally expressed, whereas *AGL40* is biparentally expressed. Maternal siRNAs mediate silencing of the maternal *AGL91* allele prior to fertilization as well as silencing of the paternal *AGL91* allele post fertilization in later stages of development. Moreover, disrupting or overexpressing *AGL91* and *AGL40* alters seed size, providing evidence for *AGLs* in endosperm development and seed size.

INTRODUCTION

Crop seeds provide nearly 70% or more calories consumed by the human population (Borlaug, 1973). Each seed consists of embryo, endosperm and seed coat. In flowering plants, embryo and endosperm result from two fertilization events: one sperm fertilizes the egg to form a diploid zygote (embryo), whereas the other sperm fertilizes the diploid central cell to form triploid endosperm. Like placenta in mammals, endosperm is the nutritional source for embryo and seed development (Moore and Haig, 1991;

[5] This chapter is being prepared for publication after initial review by: Jie Lu, Changqing Zhang, David C. Baulcombe, and Z. Jeffrey Chen.

Stebbins, 1976). The endosperm also produces 24-nt small interfering RNAs (siRNAs) that are maternally transmitted (Lu et al., 2012; Mosher et al., 2009). These maternal siRNAs are thought to protect sperm and egg from potential harmful genetic parasites including transposable elements (TEs) during sexual reproduction in plants and animals (Bourc'his and Voinnet, 2010; Martienssen, 2010; Ng et al., 2012). In animals, PiWi-interacting RNAs (piRNAs) are 25-32-nt long and mainly derived from TEs in germline (Ishizu et al., 2012). Females produce and deposit piRNAs into their eggs to confer innate immunity against potential activation of TEs in their progeny (Brennecke et al., 2007; Malone et al., 2009).

In *Arabidopsis*, biogenesis of 24-nt siRNAs requires transcription by RNA Polymerase IV (Pol IV or p4), a homologue of Pol II, generation of double-stranded RNA by RNA-DEPENDENT RNA POLYMERSE2 (RDR2), and cleavage by the endonuclease DCL3 (Haag and Pikaard, 2011; Law and Jacobsen, 2010). The resulting siRNAs are incorporated into ARGONAUTE4 (AGO4), which can target homologous transcript degradation and/or guide activities of DOMAINS REARRANGED METHYLASE2 (DRM2) (Zilberman et al., 2004) or CHROMOMETHYLASE2 (CMT2) (Zemach et al., 2013), leading to RNA-directed DNA methylation (RdDM), mainly in CHH and CHG (H = A, T, or C) sites (Haag and Pikaard, 2011; Law and Jacobsen, 2010). Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes (Ibarra et al., 2012), and DNA methylation increases in embryo after fertilization (Jullien et al., 2012), probably through actions of maternal siRNAs that are generated in endosperm (Calarco et al., 2012).

Alternatively, these siRNAs are predicted to mediate RdDM and gene expression in endosperm (Lu et al., 2012; Ng et al., 2012). Maternal siRNAs regulate expression of TE-associated genes (TAGs), including putative imprinted genes (Hsieh et al., 2011) and

those encoding Type I MADS-box transcription factors, namely, *AGAMOUS-LIKE*s (*AGL*s), which are expressed in endosperm (Belmonte et al., 2013; Bemer et al., 2010; Kradolfer et al., 2013; Lu et al., 2012). Coincidently, a recent study showed genome-wide gene expression changes among subcellular structures within the endosperm or embryo and at different stages (Belmonte et al., 2013). However, mechanisms for spatiotemporal regulation of these differentially expressed genes in endosperm, as well as the link between maternal siRNAs and differentially methylated regions in embryo and endosperm (Gehring et al., 2009; Hsieh et al., 2009) are largely unknown.

In this study, we tested a hypothesis that maternal siRNAs affect RdDM and expression of euchromatic loci in endosperm. We performed a series of genomic, genetic, and cellular experiments to determine expression and inheritance of genome-wide euchromatic siRNA loci and DNA methylation patterns in endosperm, embryo, and seed. We generated stable transgenic plants that expressed *pAGL91:AGL91::GUS* or *pAGL40:AGL40::GUS* transgene and examined their spatiotemporal expression patterns. Furthermore, we investigated the effect of disrupting or overexpressing *AGL91* and *AGL40* on endosperm development and seed size. The results collectively suggest that maternal siRNAs mediate spatiotemporal regulation of RdDM and expression of euchromatic loci including *AGLs* in endosperm, and *AGL91* and *AGL40* mediate endosperm development and seed size.

RESULTS

The parent-of-origin effect on euchromatic siRNA loci that are derived from short transposable elements (TEs) in endosperm

A major class of 24-nt siRNAs in *Arabidopsis* seed is produced by Pol IV (p4) encoded by *NRPD1A*, which is named p4-siRNAs (Mosher et al., 2009). The p4-siRNAs

in heterochromatic regions were found in both embryo and endosperm that were manually dissected, except for many siRNA loci in euchromatic regions, which were present primarily in endosperm (Figure 3.1A). Some siRNAs present in seed coat were also found in endosperm probably because soft endosperm could not be completely separated from the seed coat (Figure 3.2A). Together with published results (Lu et al., 2012), the data indicate that similar p4-siRNAs are produced in endosperm and in seeds. Thus, developing seeds were used for further analysis.

To determine inheritance of heterochromatic and euchromatic p4-siRNAs in seed development, we performed high-throughput sequencing analysis of small RNAs in seeds of reciprocal crosses between the wild-type (W) and *nrpd1a* (n) mutant plants (Table 3.1). Overall, 24-nt siRNA levels were dramatically reduced (~8%) in the *nrpd1a* mutant (nXn), compared to that (~48%) in the WXW (Figure 3.2B)**.** Moreover, siRNAs were severely reduced in the nXW cross (13%, by convention, the maternal parent is listed first in a genetic cross) compared to that in the reciprocal cross WXn (38%) (Figures 3.1B, 3.2B and C). The same trend of 24-nt siRNA reduction was also observed in *rdr2*XW (rXW) and WX*rdr2* (WXr) crosses (Figure 3.3). Because 99.1% of siRNA loci were commonly lost in both *nrpd1a* and *rdr2* mutants, further sequencing analysis was performed in *nrpd1a* and its crosses, except noted otherwise.

These p4-siRNAs were differentially distributed between heterochromatic and euchromatic regions and between leaves and seeds. The p4-siRNAs were highly enriched in heterochromatic regions in seeds as in leaves (Figures 3.1A and 3.1B) (Lu et al., 2012). However, many p4-siRNA loci in euchromatic regions were abundant in developing seeds (Figure 3.1B, blue and red) but low or absent in leaves (Figure 3.1A, black). Interestingly, there is a parent-of-origin effect on these euchromatic p4-siRNAs, which were eliminated in the nXW cross, but as highly expressed in the WXn cross as in the

WXW cross (Figures 3.1B and 3.2C). Similarly, euchromatic siRNA densities were lower in the rXW than in the WXr crosses (Figures 3.3B and 3.4A). These euchromatic p4-siRNAs dependent on maternal Pol IV are named euchromatic maternal p4-siRNAs, many of which were *TAGs* including *AGLs* (Lu et al., 2012). Examples include *AGL40* (Figures 3.1B and 3.3B) and *AGL91* (Figures 3.2C and 3.4A), in which p4-siRNAs were depleted in the nXW and rXW crosses but highly abundant in the WXn, Wxr, and WXW crosses. By contrast, there is no parent-of-origin effect on heterochromatic p4-siRNAs in pericentromeric repeats, rDNA arrays, and heterochromatic knobs. TE genes located in the pericentromeric regions such as the *Gypsy* superfamily *ATHILA2* had similar levels of p4-siRNAs in the reciprocal crosses (nXW vs. WXn and rXW vs. WXr) and in the WXW cross (Figures 3.1B, 3.2C, 3.3B, and 3.4A).

To test further, we identified 4,769 heterochromatic siRNA loci and 2,546 euchromatic siRNA loci (Materials and Methods). These siRNA loci in each category were further divided into two groups based on the ratio of siRNA counts between nXW and WXn (R = nxW/Wxn) or between rXW and WXr. The siRNA ratio was inversely correlated with the maternal dependency, and the threshold was set to 1:4 ($R \leq 0.25$) (Figures 3.4B and 3.4C). Among euchromatic siRNA loci, 593 were maternally biased ($R \leq 0.25$), and 1,953 were non-maternally biased ($R > 0.25$). Similarly, among heterochromatic siRNA loci, 401 were maternally biased ($R \leq 0.25$), and 4,368 were non-maternally biased ($r > 0.25$). Clearly, a larger proportion of maternally biased siRNAs (23.3%) was found in the euchromatic loci than in the heterochromatic loci (8.4%) ($P \approx 0$, χ^2 test). Compared to all genes, these euchromatic maternal siRNA loci were associated with TEs or TE fragments of small size (< 1-kb) (Figures 3.1C and 3.5A, left) and close to the genes; most were embedded in coding sequences (Figures 3.1C and 3.5A, right). These maternal p4-siRNA loci were enriched in RathE1-3_con TE families

(P<0.001, hypergeometric test), which were short interspersed nuclear elements (SINEs) and cut-and-paste DNA transposons (Figures 3.5B and 3.5C) and present primarily in euchromatic regions (Lenoir et al., 2001). By contrast, heterochromatic siRNA loci were mainly derived from centromeric repeats including long terminal repeat (LTR) retrotransposons such as *GYPSY* and *COPIA*.

Euchromatic maternal siRNAs correlate with RdDM in seed

The composition, location, and inheritance of these euchromatic maternal siRNAs led us to test their contributions to the RdDM in seed. We compared p4-siRNA loci with differentially methylated regions (DMRs) in developing seeds. The DMRs included hypermethylated regions in embryo (embryo DMRs) and endosperm (endosperm DMRs) at all possible cytosine methylation sites (CG, CHG and CHH, P <0.01) (Hsieh et al., 2009). The majority of CG (36,534), CHG (5,200), and CHH (8,760) sites were hypermethylated in embryo DMRs, whereas fewer CG (215), CHG (494) and CHH (989) sites were hypermethylated in endosperm (Figure 3.6A). However, the proportion of CHH methylation in endosperm DMRs (58%) was significantly higher than that in embryo DMRs (17%) ($P \approx 0$, χ^2 test), suggesting a role for RdDM in endosperm. DMRs in CHH context contained significantly more p4-siRNA loci (41.2%) than those in CG (11.6%) and CHG (31.8%), consistent with the role of siRNAs in CHH methylation ($P \approx 0$, χ^2 test) (Figure 3.6A) (Law and Jacobsen, 2010). Interestingly, siRNA loci were enriched in either endosperm or embryo DMRs, depending on maternal PolIV and RDR2 (Figures 3.6B and 3.7A). The euchromatic maternal siRNAs were significantly enriched in endosperm DMRs (14.7%) compared to all siRNA loci (3.5%) ($P \approx 0$, χ^2 test), whereas heterochromatic non-maternal siRNAs were enriched in embryo DMRs (61.1%), relative to all siRNA loci (51.4%) ($P \approx 0$, χ^2 test) (Figure 3.7B). These data depicted

two patterns of RdDM in developing seed: euchromatic maternal p4-siRNAs guided *de novo* CHH methylation of cognate loci in endosperm (Figure 3.7C, left), while heterochromatic p4-siRNAs led to DNA methylation and silencing of TEs in embryo (Figure 3.7C, right).

Among euchromatic loci, 325 genes were associated with both maternal siRNAs ($R \leq 0.25$) and endosperm DMRs. These genes were enriched in gene ontology groups of transcription factor, protein binding, and hydrolase activities ($P < 0.01$, Figure 3.8A). They included *AGLs* such as *AGL40*, *AGL85*, *AGL86*, and *AGL91*, chromatin genes such as *SET DOMAIN PROTEIN16* (*SDG16*), *SDG17*, and *SDG38*, and genes with unknown functions. These genes were in endosperm DMRs (Figure 3.8B), and their expression levels were negatively correlated with abundance of maternal siRNAs in endosperm (Lu et al., 2012).

Methylation levels of these loci (e.g., *AGL40*) were related to RdDM and reduced in the siRNA biogenesis mutants, including *nrpd1a*, *rdr2* and *dcl3*, and in the *de novo* methyltransferase mutant *drm1/2*, but not in the maintenance DNA methyltransferase mutants, *met1* and *cmt3* (Figure 3.9A). *AGL36* methylation levels were reduced in the *nrpd1a*, *rdr2*, and *cmt3* mutants but not in the *drm1/2* mutant (Figure 3.9B). Moreover, methylation levels of many loci tested were reduced in F1 crosses when *nrpd1a*, *rdr2* or *dcl3* was used as a maternal parent (Figures 3.9C-3.9F). By contrast, there was no obvious parent-of-origin effect of heterochromatic non-maternal p4-siRNAs on DNA methylation. The data suggest that euchromatic maternal p4-siRNAs guide DNA methylation of the genes in endosperm.

Spatial and temporal regulation of euchromatic siRNA genes and RdDM in endosperm

To test a role for RdDM in gene silencing, we tested expression of euchromatic siRNA target genes in embryo, seed coat, and different subregions of endosperm including peripheral (PE), micropylar endosperm (ME), and chalazal endosperm (CE) (Belmonte et al., 2013). Surprisingly, 325 target genes with both maternal siRNAs and endosperm DMRs were expressed at higher levels in CE than in other subregions (median z-score = 1.10) (Figure 3.10A, upper panel), whereas 663 target genes without endosperm DMRs showed relatively equal expression levels across all tissues tested (median z-score = 0.06, $P < 0.01$, *t* test) (Figure 3.10A, lower panel). Higher expression levels of euchromatic siRNA targets in CE are probably because they are hypomethylated. Indeed, the majority of RdDM pathway genes were repressed in CE but highly expressed in embryo or other subregions of endosperm at the heart stage (Figure 3.10B), suggesting that the RdDM pathway is inactive in CE at this stage. Within CE, RdDM pathway genes were expressed at low levels in early stages (pre-globular, globular, and heart) but at high levels in late stages (linear cotyledon and mature) (Figure 3.10C). Thus, spatiotemporal expression of siRNA target genes is anti-correlated with that of the RdDM pathway genes in endosperm.

To reveal detailed spatiotemporal expression patterns of euchromatic maternal siRNA target genes, we used a GUS (encoding β-glucuronidase) reporter of translational fusion with AGL91, which is driven by the *AGL91* promoter (Figure 3.10D). *AGL91* contained high abundance of maternal siRNAs and DMRs in endosperm but not in embryo, which is related to endosperm methylation of *AGL91* in the stage of seeds tested (Hsieh et al., 2009). These maternal siRNAs were derived from the coding sequence corresponding to the endosperm DMRs (Figures 3.7C and 3.10D). Consistent with spatial

expression of maternal siRNA target genes (Figure 3.10A), *AGL91::GUS* was localized only in the chalazal endosperm (Figure 3.10E, middle row). In contrast to the repression of RdDM in CE (Figure 3.10C), *AGL91::GUS* was expressed at pre-globular (1-2 DAP), globular (3 DAP), and heart stages (4 DAP) and silenced at the linear cotyledon stage (5-7 DAP) (Figure 3.10E, middle row). These patterns were anti-correlated with the spatiotemporal regulation of RdDM in endosperm (Figures 3.10B and 3.10C). Furthermore, in the *nrpd1a* mutant, *AGL91::GUS* expression was spread outside of the chalazal endosperm and prolonged after the linear cotyledon stage up to 7 DAP (Figure 3.10E, bottom row). The data suggest that maternal siRNAs are required for both spatial and temporal silencing of their cognate target genes.

Interestingly, *AGL91* is paternally expressed (Figure 3.11A, first row). In the cross using the wild-type (W) as a maternal parent, paternal *AGL91::GUS* expression was restricted in CE until heart stage (4 DAP). In the reciprocal cross using WT as the paternal parent, maternal *AGL91::GUS* expression was undetectable at all stages tested (Figure 3.11A, second row). Temporal silencing of *AL91::GUS* post fertilization is dependent on maternal siRNAs. In the crosses between a paternal *pAGL91:AGL91::GUS* line and a maternal mutant in RdDM (*nrpd1a*, *rdr2*, or *drm1/2*), where DNA methylation levels were reduced (Figure 3.9), *AGL91::GUS* expression prolonged after the linear cotyledon stage (5-6 DAP) but was still restricted in CE (Figures 3.11B and 3.12A). The maternal *AGL91::GUS* was not reactivated in the crosses using *nrpd1a* or *rdr2* as a paternal parent (Figure 3.11A, third row). However, when pollinating the *nrpd1a* mutant that expressed *pAGL91:AGL91::GUS* with WT pollen, maternal *AGL91::GUS* was reactivated at all stages tested and even expressed beyond CE at linear stage (Figure 3.11A, last row), suggesting that maternal *AGL91* allele is silenced by p4-siRNAs. Since RdDM pathway remains inactive in CE after fertilization until linear stage (Figures 3.10B

and 3.10C), silencing of maternal *AGL91* allele is most likely to occur before fertilization. These data suggest roles for maternal p4-siRNAs in two silencing events: spatial silencing of maternal *AGL91* allele that is likely established during female gametogenesis before fertilization and temporal silencing of paternal *AGL91* that is established post fertilization.

AGL40, another target gene with euchromatic siRNAs, also showed similar spatiotemporal expression as did *AGL91* (Figure 3.13A). However, *AGL40::GUS* did not show parent-of-origin expression patterns (Figure 3.13B). Loss of maternal siRNAs prolonged the expression of *AGL40::GUS* up to the linear cotyledon stage (6 DAP) (Figures 3.9C and 3.13C), suggesting that *AGL40* and *AGL91* has similar temporal regulation post fertilization but different allelic regulation before fertilization.

Effects for euchromatic siRNAs on seed size and embryogenesis

The effect of maternal siRNAs on *AGL91* expression is consistent with the maternal siRNAs on seed development (Lu et al., 2012). In the reciprocal crosses between wild-type (W) and *nrpd1a* (n), seeds were significantly larger in nXW than those in Wxn crosses (Figures 3.14A and 3.14B). This is related to endosperm overproliferation, which is usually accompanied by delayed embryo development (Hehenberger et al., 2012). Larger seeds showed delayed embryogenesis at 6 DAP: nxW seeds were still in the heart stage, while seeds of the WT and Wxn crosses reached the linear cotyledon stage (Figure 3.14C). Similarly, tetraploid seeds were larger than diploid seeds and also showed delayed embryogenesis (Figure 3.14C). The parent-of-origin effect on seed size is enhanced in crosses between the *nrpd1a* mutant and tetraploids (nX4 and 4Xn) (Figure 3.14C).

AGL91 and *AGL40* affect endosperm development and seed size

To test the effects of *AGL91* and *AGL40* expression on seed size, we identified T-DNA insertion lines of *AGL91* and *AGL40* (Figure 3.15A). In the insertion lines, expression of *AGL40* was reduced, and expression of *AGL91* was nearly abolished (Figure 3.15D). As a result, seed size and weight were slightly but significantly (~10%) lower in the *agl91* and *agl40* mutants than in the wild-type (Figures 3.15B and C). We also generated stable transgenic plants that expressed *AGL40* or *AGL91* under a strong endosperm-specific promoter of AT5G27880 (*SUP16*) (Wang et al., 2010) (Materials and Methods). Seed weight was significantly increased (up to ~50%) in three independently derived *AGL40* overexpression lines (Figure 3.15E and F) and positively correlated with their mRNA levels (Figure 3.15G). Seed size was also increased in the *AGL91* overexpression lines but statistically insignificant. This is probably related to redundant functions of *AGL* genes and/or spatial regulation of *AGL91*.

DISCUSSION

Pol IV-associated siRNAs maintain genome stability in embryo and spatiotemporal regulation of euchromatic loci in endosperm

In *Arabidopsis* developing seeds, p4-siRNAs are produced in endosperm and maternally transmitted (Mosher et al., 2009). Our data indicate that not all p4-siRNAs are maternally inherited. The euchromatic p4-siRNAs that are associated with short TEs exhibit the parent-of-origin effect on spatiotemporal regulation of siRNA-associated genes including *AGLs* in endosperm. However, there is no obvious maternal inheritance for the heterochromatic p4-siRNAs that are derived from long TEs in centromeric repeats and knobs, and these heterochromatic p4-siRNAs modulate epigenetic inheritance of heterochromatic loci and maintain genome stability in embryo. The maternal inheritance of euchromatic p4-siRNAs is consistent with a recent finding that the RdDM pathway is

primarily responsible for repression of euchromatic short TEs and is inhibited by heterochromatin (Zemach et al., 2013). In mammals, epigenetic modifications must be reset in the germline so their genomes undergo several rounds of DNA methylation and demethylation during differentiation of germ cells and after fertilization (Feng et al., 2010; Popp et al., 2010). In flowering plants, only CHH methylation is lost from heterochromatic long TEs in microspores and sperm cells, whereas CG and CHG methylation largely retains, and these CHH DMRs are remethylated during embryogenesis (Calarco et al., 2012). However, in endosperm, euchromatic siRNAs are maternally inherited and guide *de novo* DNA methylation during female gametogenesis and after fertilization, which establishes spatiotemporal expression patterns of these siRNA-associated genes. Activation of RdDM genes is delayed in chalazal endosperm, coincident with expression of *AGLs* prior to endosperm cellularization. Chalazal endosperm probably contains a "stem cell" niche that is enriched for the genes that are expressed during early stages of seed development, which regulates seed mass (Belmonte et al., 2013). The maternal siRNAs establish spatiotemporal regulation of DNA methylation and expression of cognate euchromatic loci including *AGLs* in endosperm, which mediates endosperm cellularization and seed size (Kang et al., 2008; Lu et al., 2012). Consequently, reducing or increasing *AGL* expression alters seed size.

Expression of paternal *AGL91* allele in chalazal endosperm is regulated by maternal siRNAs before and after zygote formation

Spatiotemporal regulation of RdDM and euchromatic loci in endosperm is consistent with the notion that the majority of imprinted genes are found in endosperm, many of which are maternally expressed (Berger and Chaudhury, 2009; Huh et al., 2007). Expression of paternal *AGL91* allele in endosperm is probably another example of paternally imprinted genes in *A. thaliana*, in addition to *PHERES1* (Kohler et al., 2005).

Our results indicate a complexity in studying imprinted genes in endosperm because imprinting in endosperm is established during female gametogenesis and temporally regulated after fertilization prior to the linear cotyledon stage. After 5 DAP, both alleles are silenced. Erasure of maternal *AGL91* silencing in the *nrpd1a* mutant alters spatial and temporal expression patterns during endosperm development. Mutations of the RdDM pathway genes including *NRPD1A*, *RDR2*, and *DRM1/2* in the maternal parent in reciprocal F1 crosses alters temporal expression of the paternal *AGL91* allele after fertilization but not the expression of the maternal *AGL91* allele that is silenced prior to fertilization. The methylation of maternal siRNA targets is restricted to endosperm instead of embryo, suggesting that this epigenetic programming of maternal alleles in endosperm is not transmitted and needs to be reset in each generation. In addition to *NRPD1a*, other factors such as DEMETER and/or Polycomb Repressive Complex 2 (PRC2) (Choi et al., 2002; Kang et al., 2008; Shirzadi et al., 2011; Zemach et al., 2013) may also coordinate silencing of maternal alleles during female gametogenesis. For example, MEDEA regulates expression of the paternal *PHE1* allele (Kohler et al., 2005). The relationships between maternal siRNAs and MEDEA and other methylation events during female gametogenesis in endosperm need further investigation.

Roles for maternal siRNAs in imprinting and seed development in flowering plants

We predict that additional imprinted genes like *AGL91* are present in chalazal endosperm prior to the linear cotyledon stage. Indeed, maternal and paternal alleles of some genes are differentially expressed (Autran et al., 2011), and paternal expression was documented (Weijers et al., 2001) during early seed development, which is likely controlled by maternal siRNAs in endosperm. If maternal p4-siRNAs in endosperm contribute to genomic imprinting as predicted (Mosher et al., 2009), eliminating these

siRNAs would alter seed development. Why is seed size not much altered in the *nrpd1a* mutant? In mammals, maternally imprinted factors inhibit growth, while paternal factors promote growth (Moore and Haig, 1991; Tilghman, 1999). As a result, the mutant phenotype of paternally imprinted *Igf2* is 40% reduction in growth (DeChiara et al., 1990), whereas the mutation of maternally imprinted *Igf2r* leads to over-growth and death (Lau et al., 1994). However, the double mutant is normal-sized and viable (Filson et al., 1993). This suggests that imprinting itself is dispensable but hijacked by fathers and mothers in the arms race to control maternal resource distributions among offspring, which is known as "the parent-offspring conflict" (Moore and Haig, 1991; Tilghman, 1999). The parent-of-origin effect of the Pol IV mutation on seed size in reciprocal crosses is consistent with this model. However, when both maternal and paternal imprints are erased in the nrpd1a mutant, there is little or no effect on the seed phenotype. NRPD1a is involved in RdDM, which establishes imprinting patterns, and the erasure of imprinting patterns affects *AGL* expression and endosperm size. This suggests an important role for RdDM in development of triploid endosperm in flowering plants. Consistent with this notion, phylogenomic studies suggest Pol IV as a major factor for the divergence between gymnosperm and angiosperm (Lee et al., 2011). Indeed, no 24-nt siRNAs are found in the gymnosperm species Pinus contorta (Morin et al., 2008), suggesting that p4-siRNAs evolve with the fertilization process in endosperm. Endosperm is more than a "yolk" or nutrient source; it has fundamental functions in embryogenesis and speciation (Costa et al., 2012). Endosperm failure is a direct cause for seed size variation or seed lethality in *Arabidopsis* interploidy crosses and interspecific hybrids (Bushell et al., 2003; Lu et al., 2012; Walia et al., 2009), which is related to dosage imbalance and/or sequence divergence between maternal siRNAs and paternal genes including TEs in closely related species (Ng et al., 2012). This model could be

further tested in interspecific hybrids to illuminate our understanding of the role for maternal siRNAs in seed development during the evolution of angiosperm.

MATERIALS AND METHODS

Plant materials and growth conditions

All plants were grown under an illumination cycle of 16-h day and 8-h night at 22^{o}C (day) and 20^{o}C (night). Flowers in the floral stage 12C (Christensen et al., 1997) were manually emasculated and pollinated at 24 hours after emasculation. Seeds were collected at designated days after pollination (DAP) for RNA extraction, GUS staining and functional assays. Wild-type (Col-0) seeds at the linear cotyledon stage (6 DAP) were manually dissected into embryo, endosperm and seed coat. Embryos were rinsed three times with 0.3 M Sorbitol before RNA extraction (Perry and Wang, 2003). Unless noted otherwise, three biological replicates were used for RNA and DNA analysis.

RNA extraction, cDNA synthesis and qRT-PCR analysis

Total RNA was extracted using Plant RNA Reagent (Invitrogen, http://www.invitrogen.com/) and treated with RQ1 DNase (Promega, http://www.promega.com/) according to the manufacturers' instructions. For cDNA synthesis, 0.5 μg of total RNA was reverse-transcribed using Superscript III reverse transcriptase (Invitrogen) and oligo(dT) primer. The cDNA was treated with RNase H (Invitrogen) at 37^{o}C for 20 min. The first-strand cDNA was used for quantitative PCR in a 15 μl-reaction using FastStart Universal SYBR Green Master Mix (Roche, http://www.roche.com) in an Applied Biosystems 7500 real-time PCR system (http://www.appliedbiosystems.com/). Quantitative PCR assays, controls and data analysis were performed as previously described (Lu et al., 2012).

Small RNA library preparation and sequencing

Small RNA libraries were prepared as previously described (Lu et al., 2012). In brief, an aliquot of 10 μg of total RNA was resolved in a 15% urea-polyacrylamide gel, and the fraction of 18-30-nt small RNAs was recovered. For *nrpd1a* and *rdr2* mutants and their crosses, whole seeds dissected from ~20 siliques were used for each library. For manually dissected seed, the embryo, endosperm and seed coat dissected from ~50 siliques were used for each library. Purified small RNAs were ligated to 5' and 3' RNA oligo adapters and reverse transcribed to produce first-strand cDNAs. PCR-amplified cDNAs were sequenced using Illumina HiSeq2000.

Analysis of small RNAs

Small RNA sequences were trimmed for 3' adapters, collapsed into unique sequences and mapped to *A. thaliana* genome (TAIR10, November 2010 release) using CASHX with perfect match option (http://carringtonlab.org/resources/cashx) (Fahlgren et al., 2009) and according to the previous protocol (Lu et al., 2012). The sequences from chloroplast, mitochondrial, rRNA, tRNAs, snoRNAs and snRNAs were excluded from the analysis. Small RNA reads were normalized by dividing the total number of reads of a library by 10 millions. Multiple-hit read was assigned equally to each locus and divided by the number of hits in the genome. siRNA loci were identified using a Python script as regions containing at least 20 distinct reads, each < 200-nt apart. The siRNA loci were then combined to a merged locus set based on their ranges in each library. A total of 9416 loci were identified. Among them, 24-nt siRNAs from 2745 (~29%) loci were not expressed in the *nrpd1a*X*nrpd1a* (nXn) cross, and 4,570 (~49%) were present in the nXn cross at a frequency <25% of Wild-typeXWild-type (WXW) (Figure S4A). Of the remaining 2,101 loci, 394 (~4%) were present in the nXn dataset with a frequency between 25-100% of WXW, and 1,707 (~18%) were close to that of WXW or greater

than WXW, many of which were siRNA loci with low abundance. Further analysis was focused on the siRNA loci that were present in the nXn dataset at a frequency <25% of WXW, which indicates Pol IV dependency. The total number for this type of loci was reduced to 7,315 (Table 3.1). Euchromatic and heterochromatic regions were defined as in previous studies (Mosher et al., 2009; Mosher et al., 2008). Small RNA distribution was displayed in Integrative Genomics Viewer (IGV) (http://www.broadinstitute.org/igv/) (Robinson et al., 2011; Thorvaldsdottir et al., 2012).

To test the statistical significance of overlaps between siRNA loci and genomic features, we randomized the genomic coordinates of the loci (keeping the length constant) 500 times. Z-scores were calculated as $z = (o - a)/s$ where o is the observed percentage of siRNA loci overlapping a given set of genomic loci, a is the average and s is the standard deviation of percentages from randomized loci (Mosher et al., 2008). The genomic locations of TEs were extracted from TAIR10 annotated TEs (Buisine et al., 2008). The distance of a TE to a protein-coding gene is calculated as the length of the sequence between a TE and its nearest protein-coding gene.

Plasmid construction

To generate plasmid *pAGL91:AGL91::GUS:3'TE*, four steps were taken. (1) A 2149-bp sequence, comprising a multiple cloning site (with 6 unique restriction sites), SV40 nuclear localization signal, restriction sites of *Pml*I and *Aat*II, and GUS coding sequence, was PCR-amplified using the plasmid DNA pFGUS2a (GenBank accession: KC920577) and synthetic oligos as templates. The amplicon was digested and inserted into the *Eco*RI/*Bam*HI of the plasmid vector pFAMIR (provided by Ramin Yadegari at University of Arizona). (2) A 964-bp sequence immediately downstream of the coding sequence of *AGL91* (At3g66656) was amplified from genomic DNA (Col-0). The

sequence was digested and inserted into the *Bam*HI/*Xma*I site of the above plasmid. (3) The 537-bp coding sequence of *AGL91* was amplified from genomic DNA (Col-0). The sequence was digested and inserted into the *Pml*I/*Aat*II site of the same plasmid. (4) The 2103-bp sequence immediately upstream of the coding sequence of *AGL91* was amplified from genomic DNA (Col-0). The sequence was digested and inserted into the *Dra*III/*Rsr*II site of the above plasmid. The sequences of primer pairs and oligo templates were listed in Table 3.2.

To generate plasmid *pAGL40:AGL40::GUS*, a 2473-bp sequence, comprising 1419-bp upstream regulatory sequence, 1041-bp coding region of *AGL40* and 13-bp linker sequence, was amplified from genomic DNA (Col-0). The sequence was digested and inserted into the *XhoI*/*Nco*I site of the plasmid vector pFGUS2a. The primer sequences were listed in Table 3.3.

To generate the overexpression constructs, the 2059-bp regulatory sequence of *SUP16* (*At5g27880*) (Wang et al., 2010) was amplified from genomic DNA (Col-0). The amplicon was digested and inserted into the *Eco*RI/*Rsr*II site of pFAMIR. The coding sequences of *AGL40* (1044-bp) and *AGL91* (537-bp) were amplified and inserted into the *Aat*II/*Nco*I site of the above plasmid. The primer pairs were listed in Table 3.4.

The above plasmid constructs were introduced into *A. thaliana* (Col-0) through the floral dip method (Bent and Clough, 1998). Transgenic lines were selected using Basta on Agrose plate. The seeds of T1 generation (T2 generation) were examined for GUS activity and seed size variation.

GUS staining and microscopy

The procedure for GUS staining was modified from a published protocol (Bemer et al., 2010). In brief, seeds were removed from siliques and stained at 37°C for 24 hours

in GUS staining solution (5 mM potassium ferricyanide, 5mM potassium ferrocyanide, 100mM sodium phosphate [pH 7.0], 0.05% Triton-X-100, 1mg mL^{-1} X-Gluc (GoldBio, https://www.goldbio.com/). Stained seeds were fixed in Carnoy's fixative (ethanol:glacial acetic acid, 3:1) for at least 3 hours, washed once with 90% ethanol, and kept in 70% ethanol for at least 24 hours. Immediately before observation, seeds were cleared in a clearing solution (chloral hydrate:glycerol:water, 8:1:2, wt/vol/vol). Pictures were taken using a compound light microscopy (Leica DM2500) equipped with Normarski Optics.

Cleared seed image

Siliques containing seeds at the desired stage were examined first under a dissecting microscope. 15 to 20 seeds were then removed from the silique and placed on the surface of a drop (1-2 cm diameter) of the clearing solution (chloral hydrate:glycerol:water, 8:1:2, wt/vol/vol) on a glass slide. After 1 hour, cleared seeds on the glass slide are visualized using a compound light microscope equipped with Nomarski optics (Leica DM2500).

Seed size and weight measurement

Seed length and width was measured using Image J software (http://rsbweb.nih.gov/ij/) on the seed images taken at 6 DAP when endosperm cellularizes and seed size is fixed (Meinke, 1994). Three biological replicates were measured each containing 50 seed images. Seed weight was measured by weighing dried mature seeds on an analytical balance (Mettler Toledo AB54-S). Three biological replicates were weighed each containing ~500 seeds. The weight was converted to mg per 100 seeds.

Analysis of DNA methylation using methylation-sensitive restriction enzyme PCR

Genomic DNA was extracted from whole seeds removed from siliques at 6 DAP using DNeasy Plant Mini kit (QIAGEN) and was digested using *Msp*I, *Hpa*II or *Hae*III at 37°C overnight. The level of methylation was measured using gel-based PCR for homozygous mutants or qPCR for crosses between mutants and the wild-type. Undigested genomic DNA was served as control and relative methylation level was normalized to wild-type selfing seeds. Primer pairs are listed in Table 3.5.

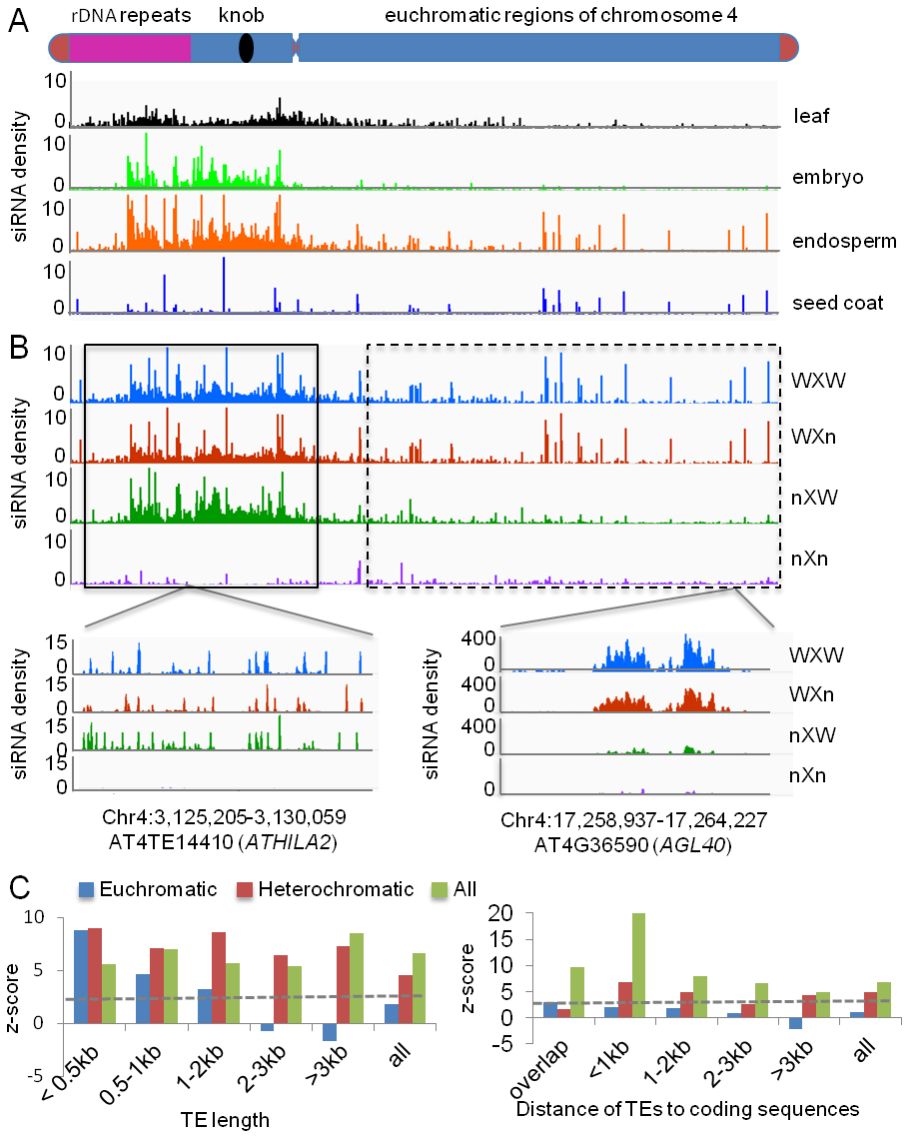

Figure 3.1: Euchromatic siRNAs are derived from short TEs and mainly in endosperm. (A) Distribution of 24-nt siRNAs on chromosome 4 in leaf (black), embryo (green), endosperm (orange), and seed coat (blue). Parts of the seed were manually dissected in the torpedo stage. (B) Chromosomal view of 24-nt siRNA distributions on chromosome 4 in seeds of WXW (WXW, blue), WX*nrpd1a* (WXn, red), *nrpd1a*XW (nXW, green), and *nrpd1a*X*nrpd1a* (nXn, purple); boxes with dashed and solid lines indicate euchromatic regions and heterochromatic regions including pericentrimeric sequences, rDNA repeats, and the knob, respectively. Inserts were examples of an euchromatic locus (*AGL40*) and a heterochromatic locus (*ATHILA2*). (C) Enrichment of 24-nt siRNAs in TEs of different size (left) and distance to gene (right) at euchromatic maternal siRNA loci (blue), heterochromatic non-maternal loci (red), and all loci (green). Dashed line indicates the statistical significance level ($P < 0.001$).

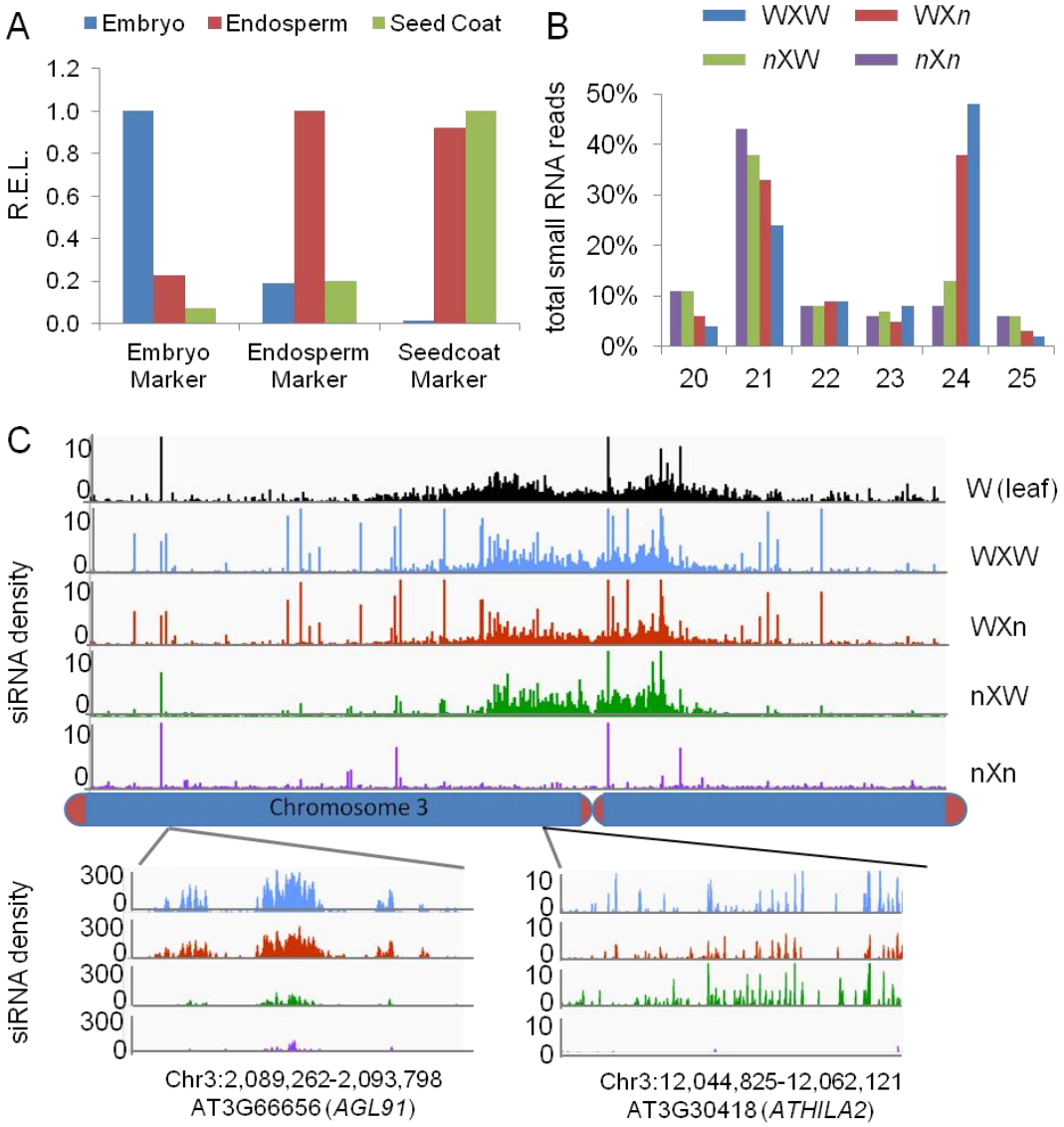

Figure 3.2: Distribution of small RNAs, and chromosomal view of siRNA distribution. (A) Relative expression levels (R.E.L.) of embryo marker (AT3G15720), endosperm marker (AT3G66656; *AGL91*), and seed coat marker (AT1G72260; *THI2.1*) genes in dissected embryo, endosperm and seed coat in the torpedo stage (6 DAP). The marker genes are selected based on the microarray data (Belmonte et al., 2013). (B) Size distribution of 20-25-nt small RNA reads in Col-0 seeds of WXW (blue), WXn (red), nXW (green) and nXn (purple) crosses. W: Wild-type; n: *nrpd1a* mutant. (C) Chromosomal view of 24-nt siRNA distributions on chromosome 3 in leaves of W (black) and in seeds of WXW (blue), WXn (red), nXW (green), and nXn (purple) crosses. Insets were examples of an euchromatic locus (*AGL91*) and a heterochromatic locus (*ATHILA2*).

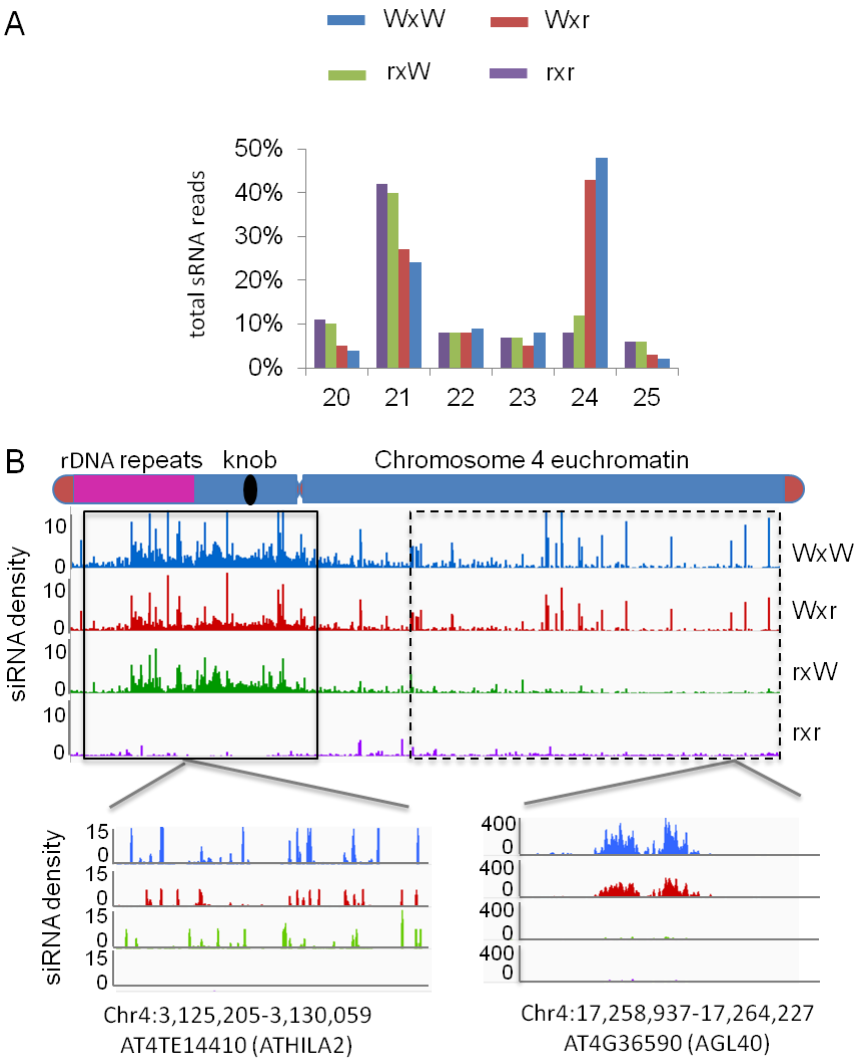

Figure 3.3: Maternal dependence of euchromatic small RNA (sRNA) in developing seeds. (A) Size distribution of 20-25-nt small RNA reads in Col-0 seeds of WXW blue), Wxr (red), rXW (green), and rxr (purple) crosses. W: Wild-type; r: *rdr2* mutant. (B) Chromosomal view of 24-nt siRNA distribution on chromosome 4 in seeds of WXW (blue), Wxr (red), rxW (green), and rxr (purple). Insets were examples of an euchromatic locus (*AGL40*) and a heterochromatic locus (*ATHILA2*).

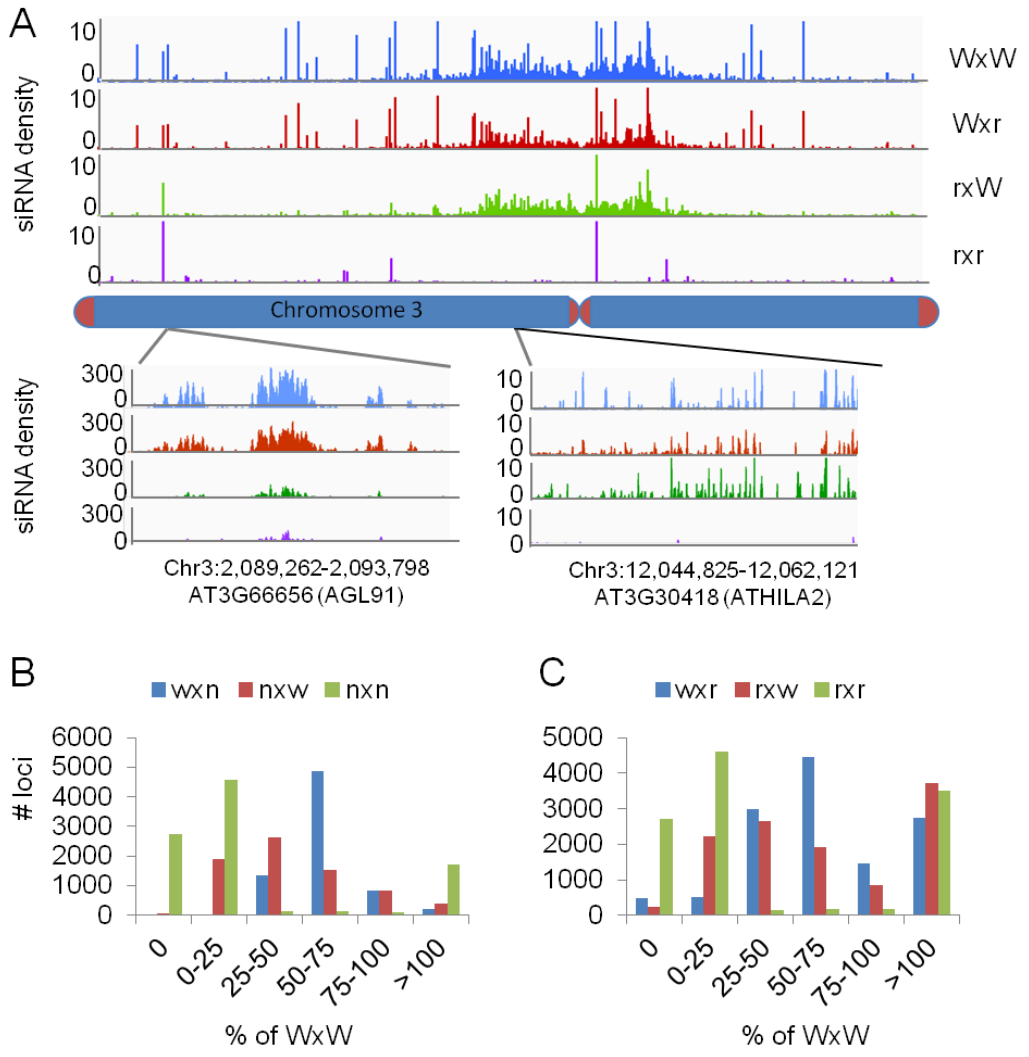

Figure 3.4: Distributions of siRNAs in F1 crosses involving *rdr2* and *nrpd1a* mutants. (A) Chromosomal view of 24-nt sRNA distribution on chromosome 3 in seeds of WXW (blue), WXr (red), rXW (green) and rXr (purple) crosses. W: wild-type; r: *rdr2* mutant. (B) Relative abundance of siRNAs compared to WXW from loci in seeds of WXn, nXW, and nXn. n: *nrpd1a* mutant. (C) Relative abundance of siRNAs compared to WXW from loci in the seeds of WXr, rXW, and rXr.

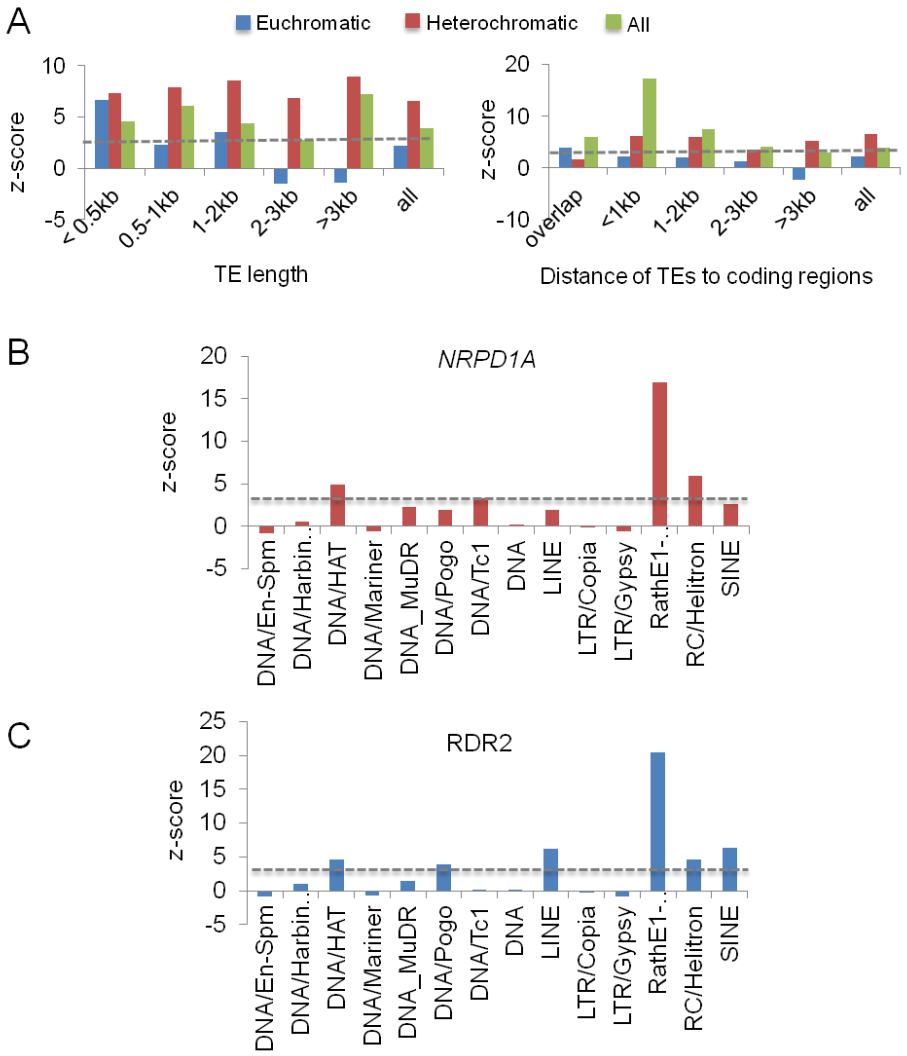

Figure 3.5: Enrichment of 24-nt siRNAs in TE families of different size and distance to genes. (A) Enrichment of 24-nt siRNAs in TEs of different size (left) and distance to genes (right) at euchromatic maternal siRNA loci (blue), heterochromatic non-maternal loci (red) and all loci (green); dashed line indicates $P < 0.001$. (B) Enrichment of siRNA loci dependent on maternal NRPD1A in different TE families; broken line indicates $P < 0.001$ (hypergeometric test). (C) Enrichment of siRNA loci dependent on maternal RDR2 in different TE families; broken line indicates $P < 0.001$ (hypergeometric test).

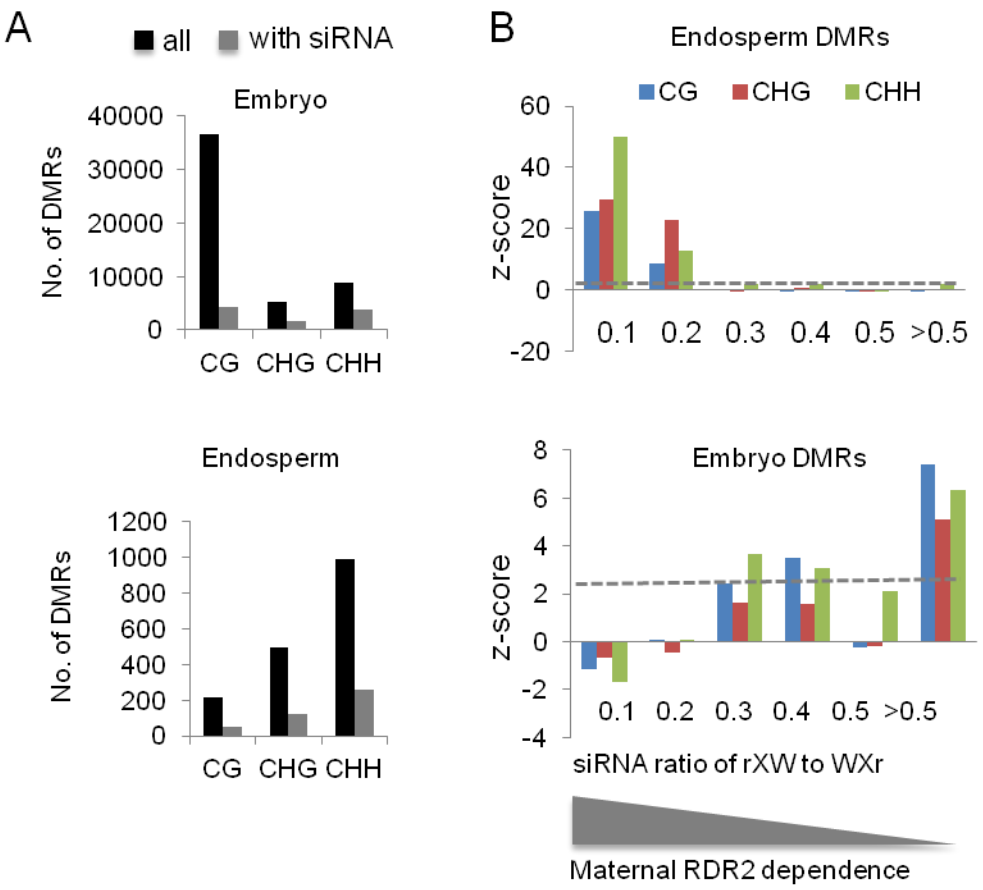

Figure 3.6: Association of maternal siRNAs with RNA-directed DNA methylation in endosperm. (A) Number of DMRs in embryo (upper) and endosperm (lower) at CG, CHG or CHH sites (Hsieh et al., 2009). (B) Proportions of endosperm DMRs (upper) and embryo DMRs (lower) at CG, CHG or CHH (H = A, T, or C) sites that are associated with RDR2-dependent siRNAs. X-axis: the ratio of siRNAs counts in rXW/WXr crosses, which is inversely correlated with the maternal RDR2 dependency. W: Wild-type; r: *rdr2* mutant.

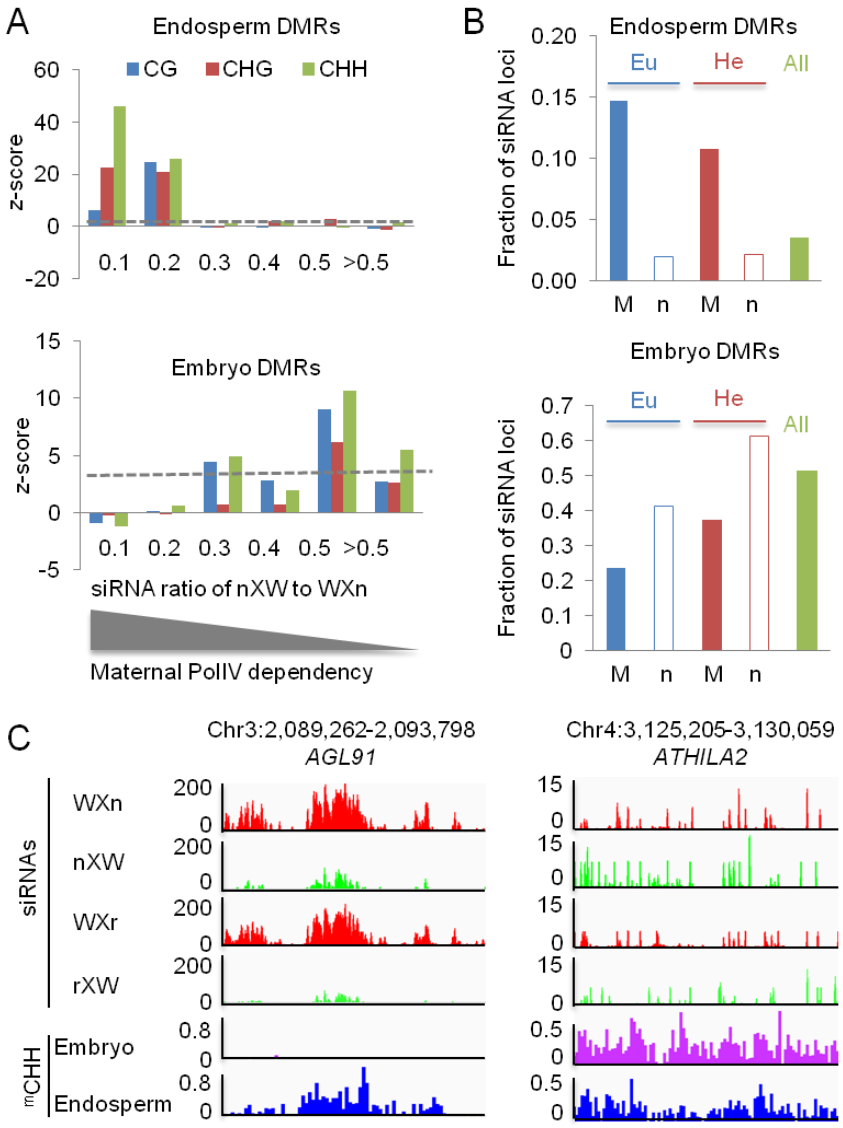

Figure 3.7: Maternal siRNAs guide RNA-directed DNA methylation in endosperm. (A) Enrichment of endosperm (En) DMRs (upper) and embryo (Em) DMRs (lower) in siRNA loci with different levels of the maternal PolIV dependency that is represented by the ratio of siRNA abundance between nxW and Wxn crosses. W: wild-type; n: *nrpd1a* mutant. (B) Fraction of siRNA loci overlapped with endosperm DMRs (upper) and embryo DMRs (lower) at maternal siRNA loci (M) or non-maternal loci (n) in euchromatic regions (Eu) or heterochromatic (He) regions. (C) Examples of an euchromatic maternal siRNA locus AT3G66656 (*AGL91*) (left) and a heterochromatic non-maternal locus AT4TE14410 (*ATHILA2*) (right) show 24-nt siRNA distributions in the seeds of Wxn (red), nxW (green), Wxr (red), and rxW (green) crosses and CHH methylation levels in embryo (purple) and endosperm (blue). r: *rdr2* mutant.

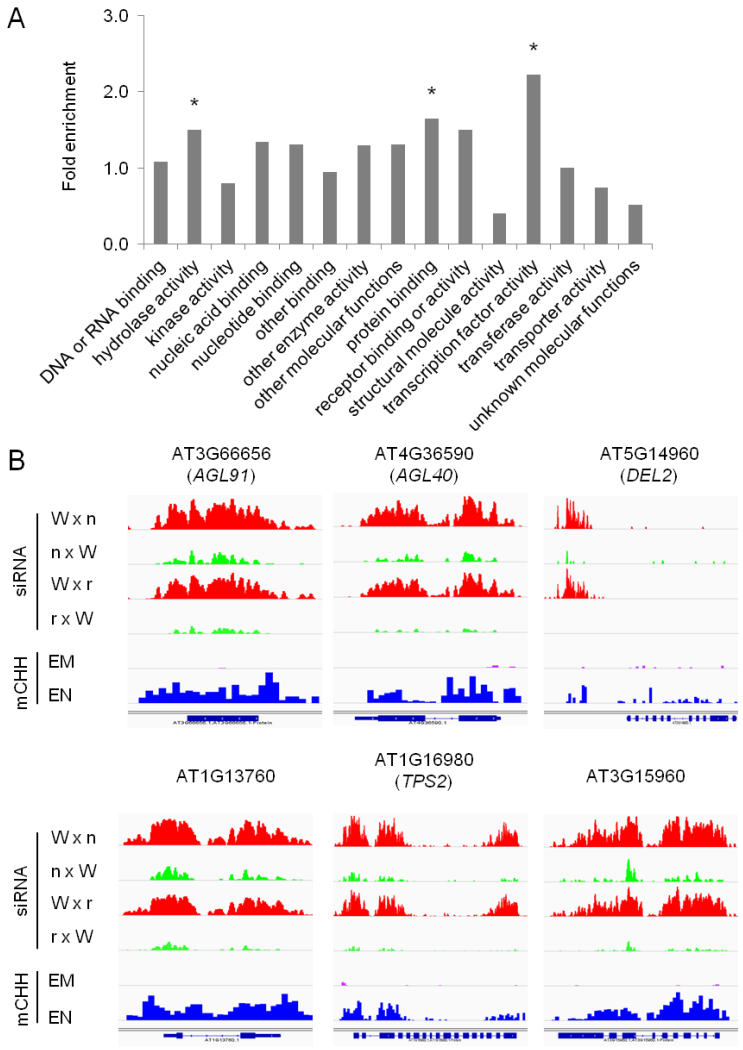

Figure 3.8: Genes associated with maternally euchromatic siRNAs and endosperm DMRs. (A) Gene ontology (GO_SLIM terms) of protein-coding genes associated with maternal euchromatic siRNAs and endosperm DMRs. Asterisks indicate FDR < 0.01 (Bonferroni multiple testing correction) using the hypergeometric test compared to the whole genome. (B) Examples of genes associated with maternal euchromatic siRNAs and endosperm DMRs. siRNA distribution is shown in red and green colors and levels of CHH methylation in endosperm (En) and embryo (Em) are shown in blue and purple colors.

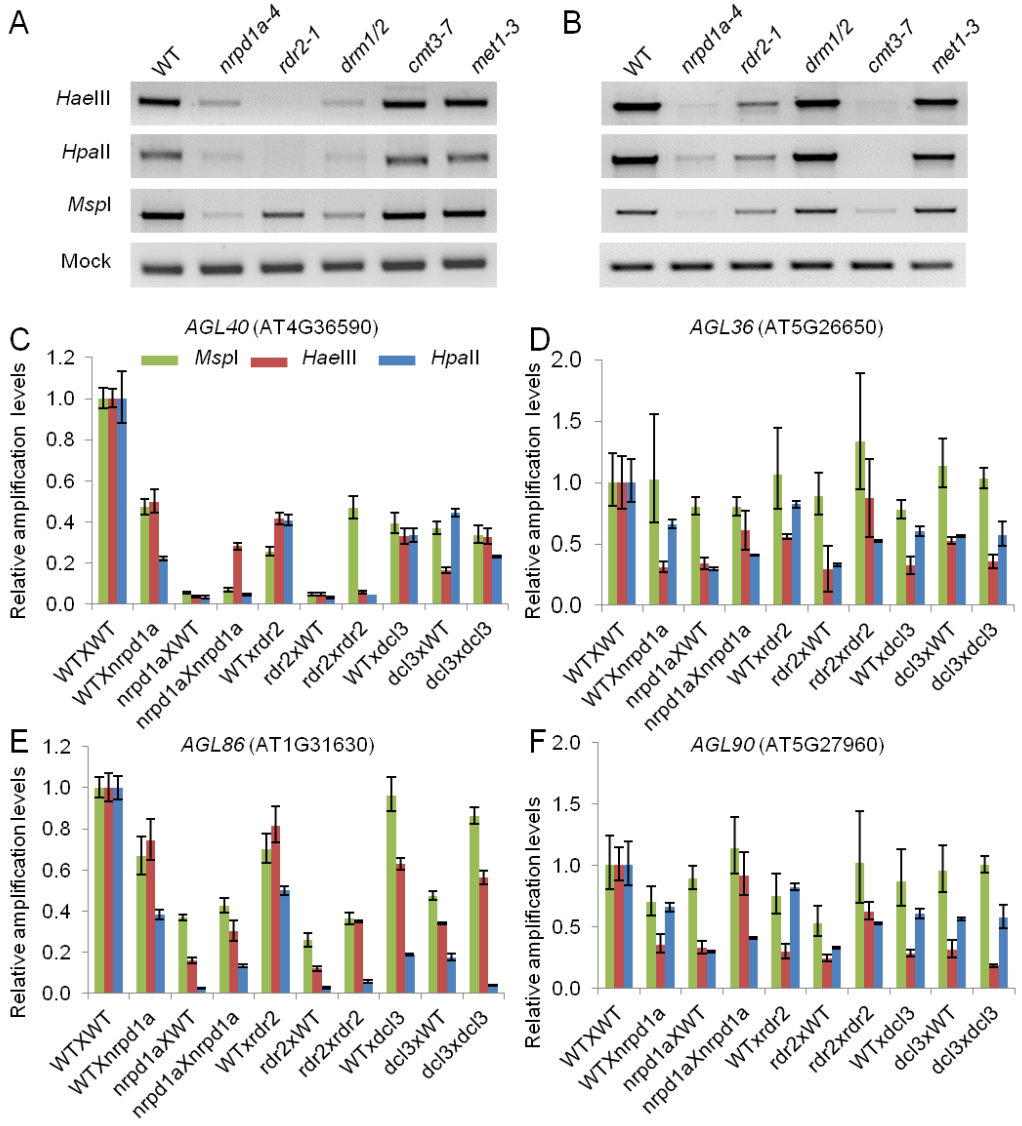

Figure 3.9: Maternal siRNA loci are associated with RdDM in endosperm and embryo. (A-B) DNA gel image analysis using methylation-sensitive restriction enzymes (MspI, HpaII, and HaeIII) of an euchromatic maternal siRNA gene (*AGL40*, AT4G36590) (A) and a heterochromatic non-maternal siRNA gene (*AGL36*, AT5G26650) (B) in the wild-type (WT), RdDM pathway gene mutants (*nrpd1a*, *rdr2*, and *dcl3*, and *drm2/1*) and maintenance methylation mutants (*met1* and *cmt3*). (C-F) qPCR analysis (relative amplification levels) using methylation-sensitive restriction enzymes (MspI, HpaII, and HaeIII) of *AGL40* (AT4G36590) (C), *AGL36* (AT5G26650) (D), *AGL86* (AT1G31630) (E), and *AGL90* (AT5G27960) (F) in seeds of reciprocal crosses between the wild-type (Col-0) and *nrpd1a*, *rdr2* and *dcl3* mutants at 6 days after pollination (DAP). Error bars = S.E.M that were derived from three independent biological replicates.

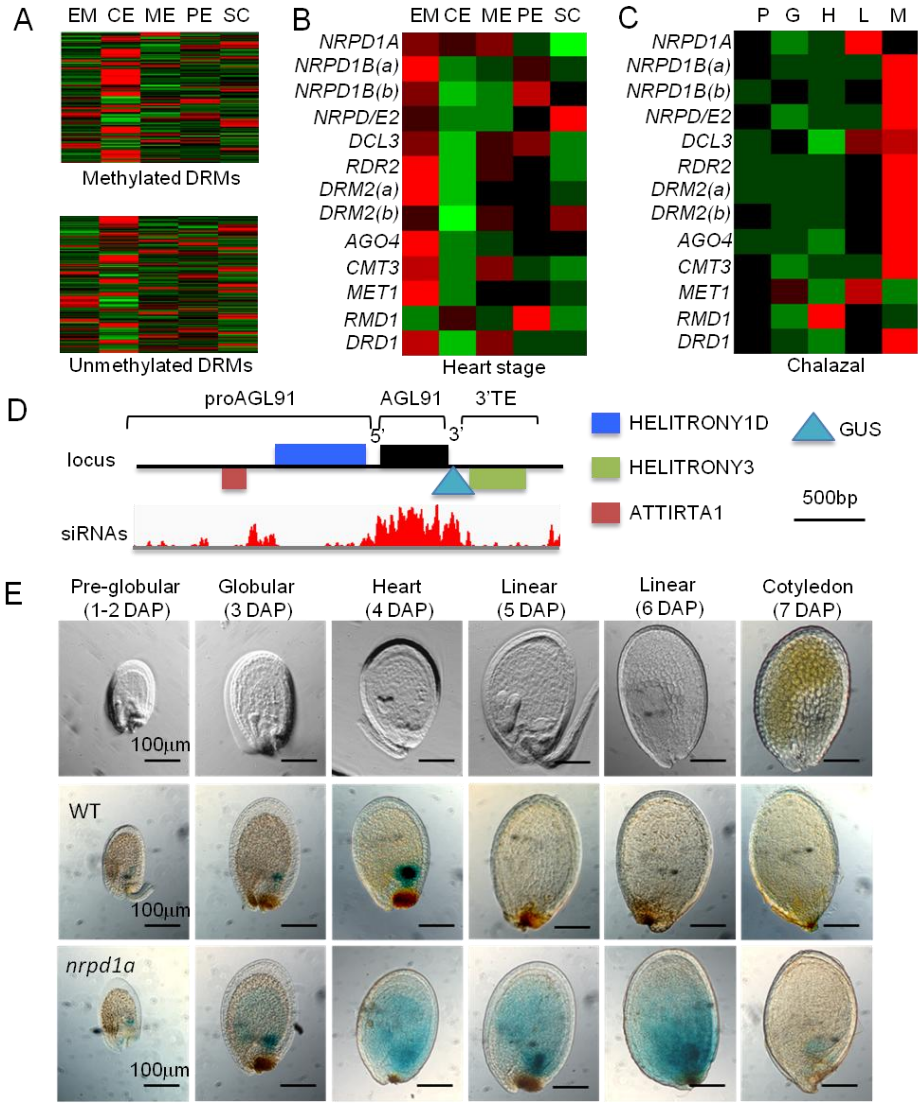

Figure 3.10: Spatiotemporal regulation of euchromatic maternal siRNA targets in endosperm. (A) Heatmap of transcript levels of the genes associated with maternal siRNAs with (upper) or without (lower) endosperm DMRs at the heart stage in embryo (EM), chalazal endosperm (CE), micropylar endosperm (ME), peripheral endosperm (PE), and seed coat (SC). (B) Heatmap of transcript levels of RdDM pathway genes in EM, CE, ME, PE, and SC at the heart stage. (C) Heatmap of transcript levels of RdDM pathway genes in CE at pre-globular (P), globular (G), heart (H), linear (L) and mature (M) stages. (D) The genomic region of *AGL91* (AT3G66656) is flanked by three short TEs, two in its promoter and one in the 3' region, and a *GUS* gene (encoding b-glucuronidase) was inserted in frame between the coding and 3' regions. (E) Seed images (upper) in different developmental stages and GUS-stained seeds of *pAGL91:AGL91-GUS-3'TE* transgenic lines in wild type (middle) or *nrpd1a* (lower). Scale bars = 0.1 mm.

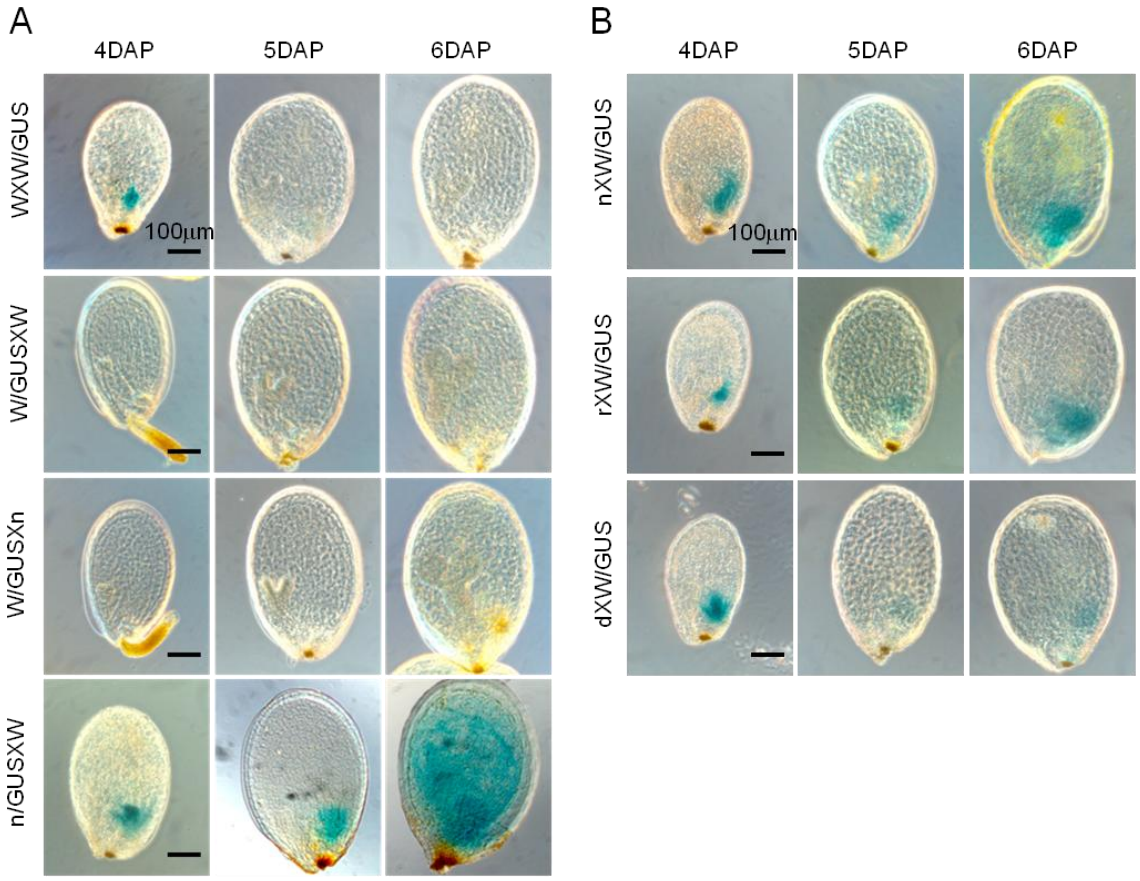

Figure 3.11: Expression of paternal *AGL91* and activation of *AGL91* in RdDM mutants. (A) GUS staining of seeds at 4-6 DAP that were isolated from WXW/GUS, W/GUSXW, and W/GUSXn and n/GUSXW crosses (n = 3x100 seeds per cross). Scale bar = 0.1 mm; W: wild-type; n: *nrpd1a* mutant; W/GUS or n/GUS: *pAGL91:AGL91:GUS* transgenic lines in wild-type or *nrpd1a* backgroud. (B) GUS staining of seeds at 4-6 DAP that were from nXW/GUS, rXW/GUS, dXW/GUS crosses (n = 3x100 seeds per cross). Scale bar = 0.1 mm; r: *rdr2* mutant; d: *drm2/1* mutant.

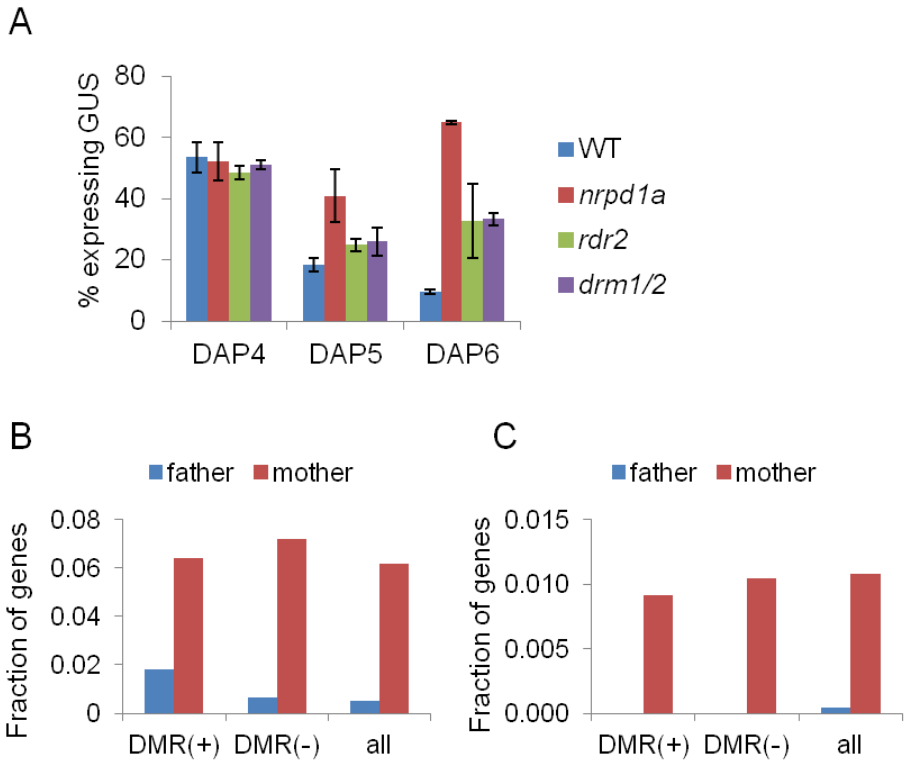

Figure 3.12: Enrichment of paternal biased genes in genes associated with maternal siRNA loci and endosperm DMRs. (A) Percentage of seeds expressing paternal GUS from WT x GUS (blue), *nrpd1a* x GUS (red), *rdr2* x GUS (green) and *drm1/2* x GUS (purple) at 4-6 DAP, error bar was derived from 3 independent biological replicates (n = 3 x100 seeds per cross). (B-C) Fraction of paternal-biased (blue) or maternal-biased (red) genes in genes associated with (+) endosperm DMRs, without (-) endosperm DMRs, and all *Arabidopsis* genes in endosperm (B) and embryo (C). Paternal- and maternal-biased genes were extracted from (Gehring et al., 2011).

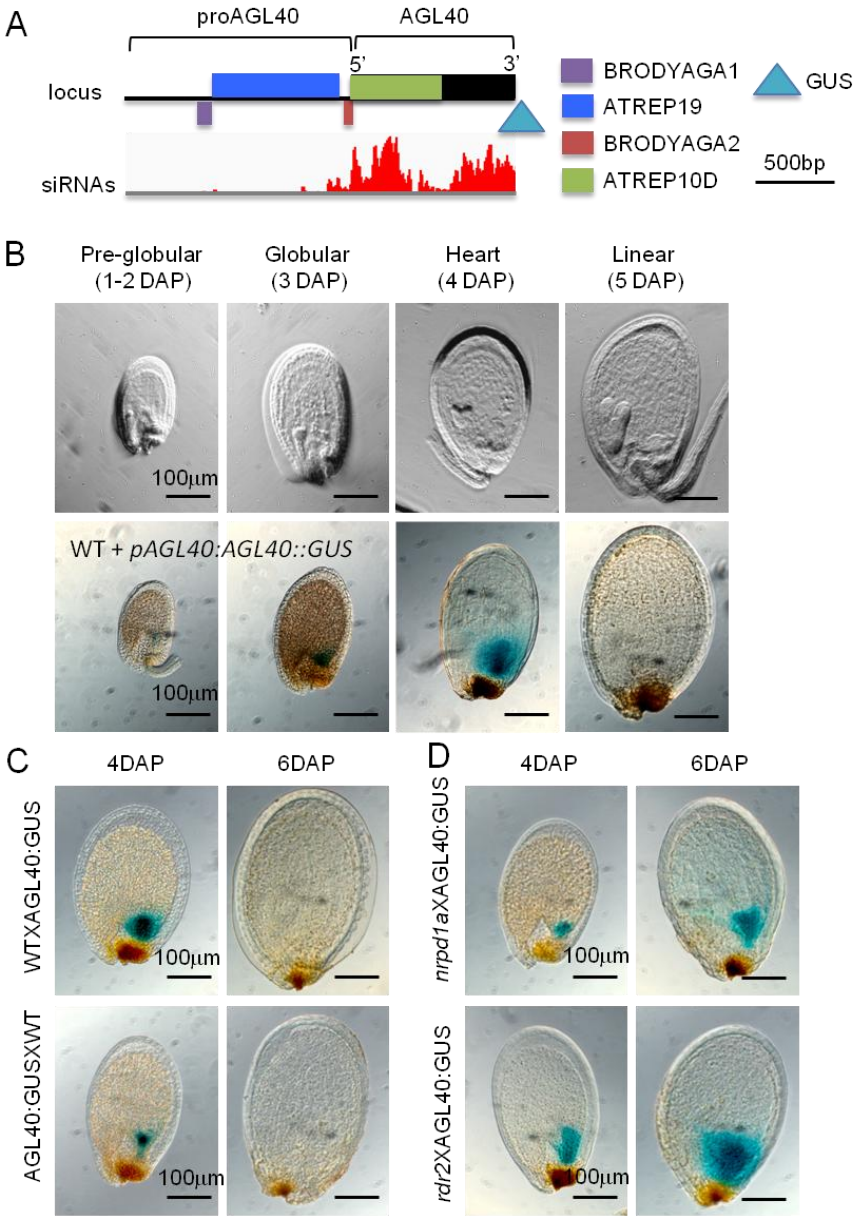

Figure 3.13: Spatial and temporal regulation of *AGL40* by maternal siRNAs in endosperm. The genomic region of *AGL40* (AT4G36590) is surrounded by four short TEs, two in its promoter and one in the coding region, and a *GUS* gene (encoding b-glucuronidase) was fused in frame after the coding region. (B) GUS-stained seeds of *pAGL40:AGL40-GUS* transgenic lines (lower) at different developmental stages (upper). (C-D) GUS-stained seeds of reciprocal crosses between *pAGL40:AGL40-GUS* transgenic lines and wild type plants (B) (n = 3 x100 seeds per cross) and seeds from and *nrpd1a* or *rdr2* x *pAGL40:AGL40-GUS* crosses (C) (n = 3 x100 seeds per cross) at 4 or 6 DAP; scale bar = 0.1mm.

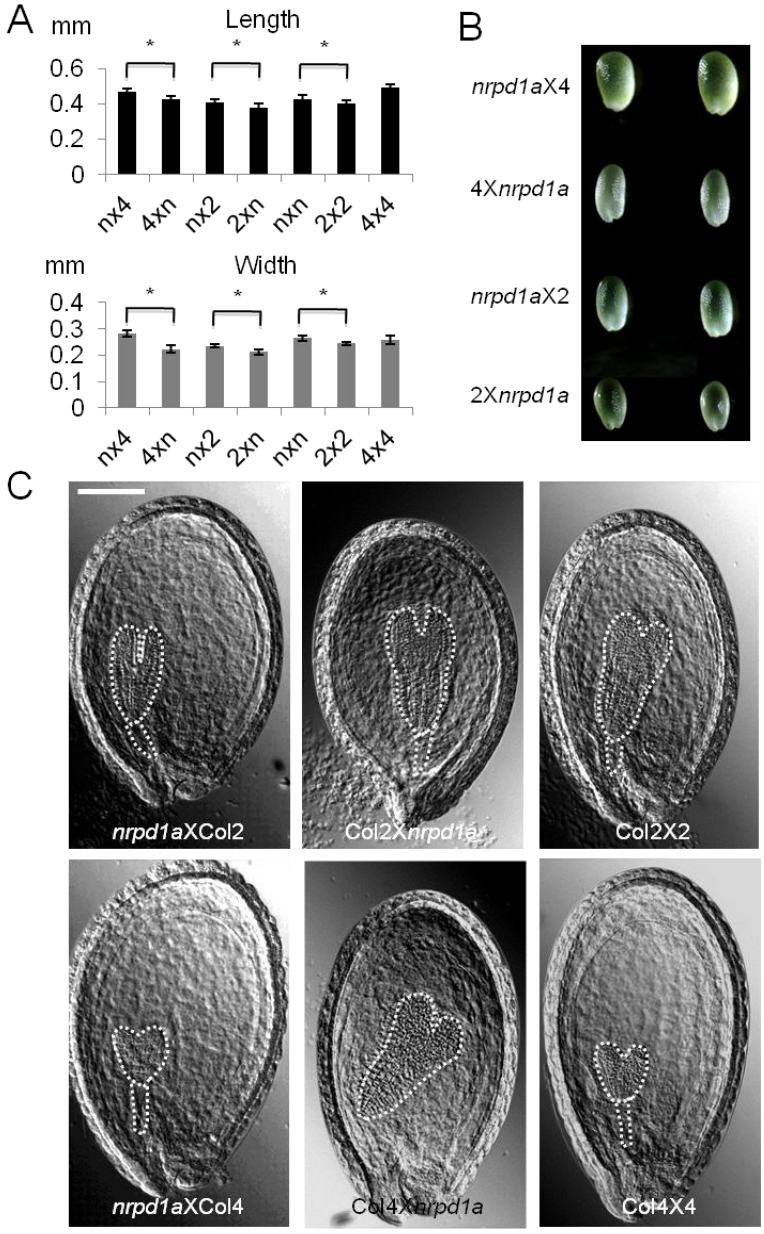

Figure 3.14: Parent-of-origin effects of p4-siRNAs on seed size and embryogenesis. (A) Length (black bar, upper) and width (grey bar, lower) of seeds from reciprocal crosses between *nrpd1a* mutant (n) and diploid (2) or tetraploid Col-0 (4) (n = 3x50 per cross); asterisk indicates $P < 0.05$ (t-test). (B) Images of developing seeds in reciprocal crosses between the *nrpd1a* mutant and a Col-0 diploid (2x) or tetraploid (4x) at 6 DAP. (C) Images of cleared seeds in crosses between Col-0 diploid (2x), tetraploid (4x), and *nrpd1a* at 6 DAP; the area corresponding to embryo is outlined with broken line; bar = 0.1 mm.

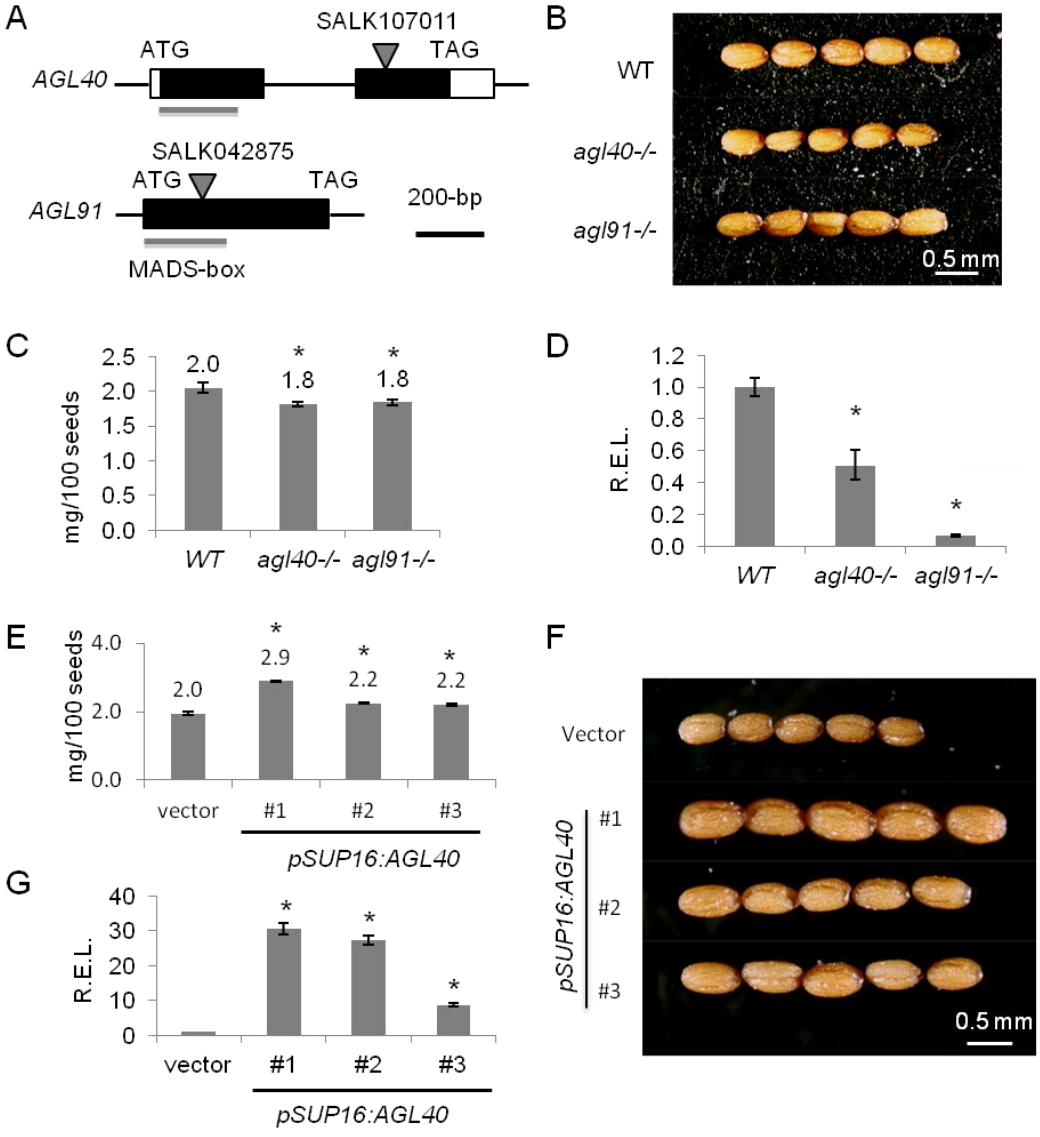

Figure 3.15: Effects of *AGL* expression on seed size. (A) Diagram of T-DNA insertion sites in *AGL40* and *AGL91*; grey bars indicate the location of MADS-box domain. (B) Images of mature seeds in T-DNA insertion lines. (C) Seed weight (mg/100 seeds) in T-DNA insertion lines of *AGL40* and *AGL91*. (D) Relative expression levels of *AGL40* and *AGL91* in T-DNA insertion lines; error bars were derived from three independent biological replicates. (E) Seed weight (mg/100 seeds) in *pSUP16:AGL40* transgenic lines (n = 3x500 seeds per line). (F) Images of mature seeds in three independent *pSUP16:AGL40* transgenic lines and the control. Scale bar = 0.5 mm. (G) Relative expression levels of *AGL40* in *pSUP16:AGL40* transgenic lines; error bars were derived from three independent biological replicates; asterisks indicate $p < 0.01$ (t-test).

Table 3.1: Summary of small RNA reads[6]

Library	**WxW**	**nxW**	**Wxn**	**rxW**	**Wxr**
total reads	10,608,788	18,869,485	21,577,253	19,467,086	22,476,947
small RNA	9,750,180	13,033,887	18,987,102	14,061,507	19,786,156
protein-coding gene	1,694,376	1,157,423	2,759,434	1,007,299	3,160,029
transposable element gene	401,864	339,714	526,687	355,883	609,335
transposable element	1,671,062	768,878	2,478,015	804,622	2,719,182
transposon fragment	1,420,384	721,331	2,057,231	758,213	2,237,277
miRNA	1,531,292	6,752,452	5,050,351	7,311,983	4,042,183
pseudogene	309,820	57,831	480,605	53,253	525,161
Library	**nxn**	**rxr**	**embryo**	**endosperm**	**seedcoat**
total reads	19,020,388	19,432,647	6,082,907	3,020,086	5,410,973
small RNA	13,493,817	13,778,133	4,468,652	2,872,324	4,128,969
protein-coding gene	1,005,858	1,072,997	367,654	505,460	477,613
transposable element gene	9,678	8,640	291,934	283,406	75,992
transposable element	176,957	173,857	695,792	867,880	401,473
transposon fragment	173,396	170,661	653,100	790,138	356,496
miRNA	7,707,381	7,865,564	94,659	224,703	297,476
pseudogene	19,333	20,252	22,159	79,113	67,852

[6] Total reads, reads perfectly match the genome; small RNA, reads after removing chloropast, mitochondrial genome, and structural noncoding RNAs, including tRNA, rRNA, snoRNA, and snRNA.

Table 3.2: The primers for generating plasmid *pAGL91:AGL91::GUS:3'TE*

	Primers
2149-bp sequence	5'- CGG AAT TCC ACA GAG TGA AGC TTC CTC AGG ATT TAA ATC GGA CCG AAC AAT GGC TCC A -3' 5'- ACG TCA GCT GCA GCC ACG TGG ACC TTT CTC TTC TTC TTT GGA GCC ATT GTT CGG TCC GA -3' 5'- CAC GTG GCT GCA GCT GAC GTC GCT GCA GCT GCT GCA GTA GAT CTG AGG GTA AAT TTC TAG -3' 5'- CGG GAT CCT CAT TGT TTG CCT CCC TGC TG -3'
964-bp sequence	5'- CGG GAT CCC AGT TAT GCA AAT GAG AAG GCA -3' 5'- TAT CTC TCT CCC GGG AAT GAA AAT TTC TCT TCC TTG CTT CCT -3'
537-bp sequence	5'- AGA GAC TGA CAC GTG ATG GGT AGG AGA AAG ATT AAG ATG GA -3' 5'- AGA GGA GAG GAC GTC ATT ATC ATT AGA GAG AAA CAT GAG AGA AG -3'
2103-bp sequence	5'- ACT GAC AGT CAC AGA GTG TGA AAG CAT TTT TCA TTA TAT ATA TAT AC -3' 5'- GAG AAG AGA CGG TCC GTT CTT TTT TTT TGT GTG AAA TGT TTT GAG -3'

Table 3.3: The primers for generating plasmid *pAGL40:AGL40::GUS*

	Primers
2473-bp sequence	5'- ATT CTC CTC CCG CTC GAG CTT GGT TCC AAT CTT CAT GGA G -3' 5'- AAC ACA GTC ATG CCA TGG CAG CAG CAG CAG CGC TCT GGT TGA AGT TGT AAC TTG AC -3'

Table 3.4: The primers for generating *AGL40* and *AGL91* overexpression lines.

	Primers
2059-bp regulatory sequence of *SUP16*	5’ – GCG AAA GAA TTC CTG AAT GTG CAA ACA AAC ATG TC – 3’ 5’ – AAA CCC CGG TCC GAG CTG GTT CTC TGT AAC AAA TC – 3’
AGL40 coding sequenc	5’ – CCG AAA GAC GTC AAC AAT GGT GAG AAG TAC CAA AGG T – 3’ 5’ – CCC TAA CCA TGG CTA GCT CTG GTT GAA GTT GTA AC – 3’
AGL91 coding sequence	5’ – AAG AAG GAC GTC AAC AAT GGG TAG GAG AAA GAT TAA GAT G – 3’ 5’ – AAA GGA CCA TGG CTA ATT ATC ATT AGA GAG AAA CAT GAG AG – 3’

Table 3.5: qPCR primers for methylation sensitive restriction enzyme assay

Gene	Primer sequences
AGL40	5’ – GGTGGGAAAGTGTTTTCTTT – 3’ 5’ – GAGTATATTGTTGAGATATTGGATATTTCT AT – 3’
AGL36	5’ – GCTTGTGCTCTCATCTACAGT – 3’ 5’ – GGCATCTCCAGAAACCTTGAA – 3’
AGL86	5’ – GTACGAGAATCCAGTGGTG – 3’ 5’ – GACTCTCCAGTTTCTTTGTTTC – 3’
AGL90	5’ – ACGGGATATTCAAGAAACTCCA – 3’ 5’ – CGGCATCTCCAGAAACTTTG – 3’

Chapter 4: Discussion and Future Directions

BIOGENESIS OF P4-SIRNAS FROM *AGL*S

We showed that production of p4-siRNAs from *AGL*s was dependent on NRPD1A and RDR2 (Figure 3.1). It is not clear how these specific sequences are recognized by the biogenesis machinery. A recent report has shown that DNA BINDING TRANSCRIPTION FACTOR 1 (DTF1) may assist in the recruitment of PolIV by binding to H3K9me1/2 and physically interacting with CLSY1(Zhang et al., 2013). However, p4-siRNAs are most abundantly expressed from the coding regions of *AGL*s (Figures 2.9 and 2.10) which are actively transcribed in developing seeds (Figure 2.7) (Bemer et al., 2010). Since H3K9me1/2 are the hallmarks of repressive heterochromatin (Pfluger and Wagner, 2007), it is unlikely that actively expressed *AGL*s are enriched in repressive marks. Alternatively, actively transcription by PolII may be required for p4-siRNA production. Indeed, *AGL*-derived p4-siRNAs are specifically expressed in developing seeds but not expressed in vegetative tissues where most Type I MADS-box genes are silent (Figure 3.1) . Therefore, it is tentative to speculate that active transcription by PolII switches the repressive chromatin status of *AGL*s in vegetative tissues to an active status in seeds, which makes *AGL*s accessible to RdDM. A recent report supports the notion that RdDM is inhibited by heterochromatin including H3K9me1/2 (Zemach et al., 2013). RdDM reinforces p4-siRNA production (Pontier et al., 2005; Zilberman et al., 2004) and it is probably through this feedback loop that the abundant levels of *AGL*-derived siRNAs are accumulated in seeds. Analysis of small RNAs in PolII mutant may be performed to test whether active transcription by PolII is required for p4-siRNA biogenesis in seeds. Chromatin immunoprecipitation (ChIP) in both vegetative tissues and developing seeds will help to determine the change of

chromatin modifications from vegetative to reproductive growth and how this change facilitates p4-siRNA production from *AGL*s.

TRANSCRIPTIONAL AND POST-TRANSCRIPTIONAL SILENCING

Many *TAG*s including Type I MADS-box genes contain extensive gene body methylation at all sequence contexts corresponding to the maternal p4-siRNA loci (Figures 3.8B). In eukaryotes, substantial methylation is found in the bodies of active genes, where methylation is generally restricted to the CG context and is not thought to lead to gene silencing (Law and Jacobsen, 2010; Zemach et al., 2010; Zilberman et al., 2007). In fact, moderately transcribed genes are most likely to be methylated in genic regions (Zilberman et al., 2007). However, we showed that p4-siRNAs-mediated DNA methylation led to transcriptional silencing of *AGL*s in a spatiotemporal manner. Our analysis was carried out in a dynamic developmental scope and the silencing of *AGL*s occurs only at a very narrow time window in specific regions. Thus it is necessary to take developmental and environmental effects under consideration when investigating the function of DNA methylation. It is possible that gene body methylation may trigger silencing in a specific tissue, during a certain developmental stage or upon a certain environmental cue.

Post-transcriptional gene silencing may also contribute to the silencing of *AGL*s. It is noteworthy that in addition to the 24-nt siRNAs, there are considerable amount of 21-nt siRNAs generated from *AGL*s (Figure 2.11E). In plants, the majority of 21-nt siRNAs are miRNAs that mediate mRNA cleavage and repress gene expression at post-transcriptional level. Therefore, the 21-nt siRNAs may coordinate *AGL* silencing through cleavage of the *AGL* transcripts. Interestingly, biogenesis of *AGL*-derived 21-nt siRNAs also depends on NRPD1A (Figure 2.11E), suggesting the generation of 21-nt siRNAs

may be routed to microRNA biogenesis pathway besides RDR2/DCL3 pathway after initial transcription by PolIV. Alternatively, a tasiRNA-like mechanism could be involved in triggering secondary siRNA cascades. Indeed, *AGL91*, *AGL40*, and *AGL36* had a significantly high probability ($P < 0.01$) of generating 21-nt phased siRNAs (Table 2.4). In *Arabidopsis*, Dicer-like 1 is responsible for generating 21-nt miRNAs from their precursors while RNA-dependent RNA Polymerase 6 converts miRNA cleaved mRNAs into double-stranded RNAs which are subsequently cut by Dicer-like 4 to 21-nt tasiRNAs (Xie et al., 2004). To test whether post-transcriptional silencing is involved in *AGL* repression, small RNA and transcriptome analysis can be performed in microRNA biogenesis as well as in tasiRNA biogenesis mutants in developing seeds.

GENOMIC IMPRINTING OF *AGL*S

AGL91 was paternally expressed while *AGL40* was biparentally expressed (Figures 3.11 and 3.13). Although *AGL91* and *AGL40* share similar patterns of maternal p4-siRNA expression and DNA methylation (Figure 3.8B), genomic imprinting only occurs at *AGL91* locus. Interestingly, the 3' downstream sequence of *AGL91* contains a TE which is absent from that of *AGL40* (Figures 2.9C and G). This is reminiscent of the case of *PHE1* and its close homolog *PHE2* (Villar et al., 2009). *PHE1* contains a tandem repeat in its 3' region and is paternally expressed while *PHE2* does not contain tandem repeat and is not regulated by genomic imprinting. The tandem repeat in the 3' region of *PHE1* is required for the imprinting (Villar et al., 2009). It would be interesting to test whether the 3' TE in *AGL91* is necessary or sufficient to for its paternal expression pattern. Furthermore, *AGL91* had a higher p4-siRNA level than *AGL40* (Table 2.4), which is likely to contribute to the overall higher transcript level of *AGL40* compared to

AGL91(Belmonte et al., 2013). Therefore, the lower level of p4-siRNAs in *AGL40* may not be sufficient to completely block the maternal allele.

Neither *AGL91* nor *AGL40* was identified in previous sequencing experiments because both were silenced at 6-7 DAP (Figures 2B and 3A), when the RNA and DNA were prepared for sequencing experiments (Gehring et al., 2009; Gehring et al., 2011; Hsieh et al., 2009). Since RdDM is active in chalazal endosperm at 5 DAP and most imprinted genes are repressed thereafter (Belmonte et al., 2013), the choice of stage is critical for discovery of new imprinted genes.

P4-SIRNAS AS SPECIES BARRIER

Barbara McClintock predicted in 1984 that TEs could be derepressed and mobilized in response to the "genomic shock" (McClintock, 1984). In the cross between *A.thaliana* diploid and its close relative *A.arenosa* tetraploid which results in high seed lethality, the normally silenced and heterochromatic element *ATHILA* was expressed from the paternal, but not maternal chromosomes (Josefsson et al., 2006).

It is tempting to speculate that sequence divergence or higher copy numbers of TEs in paternal genome can escape the suppression from the siRNAs produced by maternal genome. Recent studies on *Drosophila* hybrid dysgenesis provide strong evidence for the crucial role of small RNAs in TE mobilization during hybridization. Hybrid dysgenesis has been characterized in many *Drosophila* species (Bingham et al., 1982; Engels and Preston, 1979; Kidwell, 1981). For example, in *Drosophila melanogaster*, the progeny of crosses between wild-caught males and laboratory-strain females are sterile, whereas the genetically identical progeny of the reciprocal cross remain fertile (Kidwell, 1977; Picard, 1976). This was attributable to the mobilization of P-element or I-element transposons, which were present in wild-caught flies but absent

from laboratory strains, leading to defects in gametogenesis (Bucheton et al., 1984; Castro and Carareto, 2004; Chambeyron and Bucheton, 2005; Kidwell, 1983; Pelisson, 1981; Rubin et al., 1982). In germline cell, piRNAs epigenetically repress the mobilization of TEs and are crucial to normal gametogenesis. These piRNAs are maternally deposited into the oocytes of the daughters (Brennecke et al., 2008). Laboratory-strain females lacking P-element and I-element were not able to deposit enough piRNAs to their daughters' oocytes so that the TEs from paternal chromosomes mobilized and disrupted gametogenesis (Brennecke et al., 2008).

Heterochromatic TEs evolves much more rapidly than euchromatic genes and diverge greatly between strains, varieties and species. p4-siRNA originated from heterochromatic regions may serve as species barriers during hybrid formation. Maternal inheritance of p4-siRNAs makes the copy number and sequence divergence of TEs in paternal genome very crucial in determining the compatibility of two species in hybrids. A probable model is that p4-siRNAs are provided by the maternal alleles in central cell to the zygotes to suppress TE activities since the two maternal nuclei in the central cell undergo genome-wide demethylation which does not occur in egg cell and sperm. These siRNAs guide *de novo* methylation to the corresponding sequences and suppress their expression. When paternal genome has extra copies of TEs or TEs that are absent from maternal genome, their offspring are at stake. Chances are that TEs will be mobilized and insert into important embryogeneis genes, leading to postzygotic lethality characterized many interspecific hybrids. Collectively, small RNAs serve as links between genome wide gene expression changes and epigenetic reprogramming, which shape the physiological and morphological renovation of interspecies hybrids and allopolyploids.

Maternally inherited p4-siRNAs are most likely to account for the unleashing of TEs in response to the "genomic shock". It has yet to be determined whether maternally

expression patterns of p4-siRNAs will be maintained or disrupted. How divergent are p4-siRNA loci in closely related species? How does the divergence of p4-siRNAs affect postzygotic lethality in interspecific hybrids? Can we ameliorate the hybrid seed lethality by expressing paternal genome specific p4-siRNA loci in female gametophytes? A better understanding of small RNA regulation in interspecific hybrids and polyploids will help us effectively select the best combinations of parents for producing hybrids and polyploidy plants and manipulate small RNA expression to overcome species barriers and produce “super hybrids” for better food, feed and fuels.

References

Adams, S., Vinkenoog, R., Spielman, M., Dickinson, H.G., and Scott, R.J. (2000). Parent-of-origin effects on seed development in Arabidopsis thaliana require DNA methylation. Development *127*, 2493-2502.

Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350-355.

Autran, D., Baroux, C., Raissig, M.T., Lenormand, T., Wittig, M., Grob, S., Steimer, A., Barann, M., Klostermeier, U.C., Leblanc, O.*, et al.* (2011). Maternal epigenetic pathways control parental contributions to Arabidopsis early embryogenesis. Cell *145*, 707-719.

Axtell, M.J. (2013). Classification and Comparison of Small RNAs from Plants. Annu Rev Plant Biol.

Axtell, M.J., Snyder, J.A., and Bartell, D.P. (2007). Common functions for diverse small RNAs of land plants. Plant Cell *19*, 1750-1769.

Baroux, C., Gagliardini, V., Page, D.R., and Grossniklaus, U. (2006). Dynamic regulatory interactions of Polycomb group genes: MEDEA autoregulation is required for imprinted gene expression in Arabidopsis. Genes Dev *20*, 1081-1086.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215-233.

Baulcombe, D. (2004). RNA silencing in plants. Nature *431*, 356-363.

Belmonte, M.F., Kirkbride, R.C., Stone, S.L., Pelletier, J.M., Bui, A.Q., Yeung, E.C., Hashimoto, M., Fei, J., Harada, C.M., Munoz, M.D.*, et al.* (2013). Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. Proc Natl Acad Sci U S A *110*, E435-444.

Bemer, M., Heijmans, K., Airoldi, C., Davies, B., and Angenent, G.C. (2010). An atlas of type I MADS box gene expression during female gametophyte and seed development in Arabidopsis. Plant Physiol *154*, 287-300.

Bemer, M., Wolters-Arts, M., Grossniklaus, U., and Angenent, G.C. (2008). The MADS domain protein DIANA acts together with AGAMOUS-LIKE80 to specify the central cell in Arabidopsis ovules. Plant Cell *20*, 2088-2101.

Bent, A.F., and Clough, S.J. (1998). *Agrobacterium* germ-line transformation: Transfromation of *Arabidopsis* without tissue culture. In Plant Molecular Biology Manual, S.B. Gelvin, and R.A. Schilperoort, eds. (Dordrecht, The Netherlands, Kluwer Academic Pub.), pp. 1-14.

Berger, F. (1999). Endosperm development. Curr Opin Plant Biol *2*, 28-32.

Berger, F. (2003). Endosperm: the crossroad of seed development. Curr Opin Plant Biol *6*, 42-50.

Berger, F. (2008). Double-fertilization, from myths to reality. Sexual Plant Reproduction *21*, 3-5.

Berger, F., and Chaudhury, A. (2009). Parental memories shape seeds. Trends Plant Sci *14*, 550-556.

Bingham, P., Kidwell, M., and Rubin, G. (1982). The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. Cell *29*, 995-1004.

Borlaug, N.E. (1973). Civilization's future: a call for international granaries. Science and Public Affairs *24*, 7-15.

Bourc'his, D., and Voinnet, O. (2010). A small-RNA perspective on gametogenesis, fertilization, and early zygotic development. Science *330*, 617-622.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell *128*, 1089-1103.

Brennecke, J., Malone, C.D., Aravin, A.A., Sachidanandam, R., Stark, A., and Hannon, G.J. (2008). An Epigenetic Role for Maternally Inherited piRNAs in Transposon Silencing. Science *322*, 1387-1392.

Brown, R.C., Lemmon, B.E., Nguyen, H., and Olsen, O.-A. (1999). Development of endosperm in Arabidopsis thaliana. Sexual Plant Reproduction *12*, 32-42.

Bucheton, A., Paro, R., and Sang, H.e.a. (1984). The molecular basis of I-R hybrid dysgenesis in Drosophila melanogaster: identification, cloning, and properties of the I factor. Cell *38*, 153-163.

Buisine, N., Quesneville, H., and Colot, V. (2008). Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics *91*, 467-475.

Bushell, C., Spielman, M., and Scott, R.J. (2003). The basis of natural and artificial postzygotic hybridization barriers in Arabidopsis species. Plant Cell *15*, 1430-1442.

Calarco, J.P., Borges, F., Donoghue, M.T., Van Ex, F., Jullien, P.E., Lopes, T., Gardner, R., Berger, F., Feijo, J.A., Becker, J.D.*, et al.* (2012). Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. Cell *151*, 194-205.

Cantu, D., Vanzetti, L.S., Sumner, A., Dubcovsky, M., Matvienko, M., Distelfeld, A., Michelmore, R.W., and Dubcovsky, J. Small RNAs, DNA methylation and transposable elements in wheat. BMC Genomics *11*, 408.

Castro, J., and Carareto, C. (2004). Drosophila melanogaster P transposable elements: mechanisms of transposition and regulation. Genetica *121*, 107-118.

Chambeyron, S., and Bucheton, A. (2005). I elements in Drosophila: in vivo retrotransposition and regulation. Cytogenet Genome Res *110*, 215-222.

Chan, S.W., Zhang, X., Bernatavichute, Y.V., and Jacobsen, S.E. (2006). Two-step recruitment of RNA-directed DNA methylation to tandem repeats. PLoS Biol *4*, e363.

Chapman, E.J., and Carrington, J.C. (2007). Specialization and evolution of endogenous small RNA pathways. Nat Rev Genet *8*, 884-896.

Chaudhury, A.M., Ming, L., Miller, C., Craig, S., Dennis, E.S., and Peacock, W.J. (1997). Fertilization-independent seed development in Arabidopsis thaliana. Proc Natl Acad Sci U S A *94*, 4223-4228.

Chen, H.M., Li, Y.H., and Wu, S.H. (2007). Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis. Proc Natl Acad Sci U S A *104*, 3318-3323.

Chen, X. (2009). Small RNAs and their roles in plant development. Annu Rev Cell Dev Biol *25*, 21-44.

Chen, X. (2012). Small RNAs in development - insights from plants. Curr Opin Genet Dev *22*, 361-367.

Choi, Y., Gehring, M., Johnson, L., Hannon, M., Harada, J.J., Goldberg, R.B., Jacobsen, S.E., and Fischer, R.L. (2002). DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in arabidopsis. Cell *110*, 33-42.

Christensen, C.A., King, E.J., Jordan, J.R., and Drews, G.N. (1997). Megagametogenesis in Arabidopsis wild type and the Gf mutant. Sex Plant Reprod *10*, 49-64.

Church, G.M., and Gilbert, W. (1984). Genomic sequencing. Proc Natl Acad Sci U S A *81*, 1991-1995.

Costa, L.M., Yuan, J., Rouster, J., Paul, W., Dickinson, H., and Gutierrez-Marcos, J.F. (2012). Maternal control of nutrient allocation in plant seeds by genomic imprinting. Curr Biol *22*, 160-165.

Day, R.C., Herridge, R.P., Ambrose, B.A., and Macknight, R.C. (2008). Transcriptome analysis of proliferating Arabidopsis endosperm reveals biological implications for the control of syncytial division, cytokinin signaling, and gene expression regulation. Plant Physiol *148*, 1964-1984.

De Bodt, S., Raes, J., Florquin, K., Rombauts, S., Rouze, P., Theissen, G., and Van de Peer, Y. (2003a). Genomewide structural annotation and evolutionary analysis of the type I MADS-box genes in plants. J Mol Evol *56*, 573-586.

De Bodt, S., Raes, J., Van de Peer, Y., and Theissen, G. (2003b). And then there were many: MADS goes genomic. Trends Plant Sci *8*, 475-483.

DeChiara, T.M., Efstratiadis, A., and Robertson, E.J. (1990). A growth-deficiency phenotype in heterozygous mice carrying an insulin-like growth factor II gene disrupted by targeting. Nature *345*, 78-80.

Dilkes, B.P., Spielman, M., Weizbauer, R., Watson, B., Burkart-Waco, D., Scott, R.J., and Comai, L. (2008). The maternally expressed WRKY transcription factor TTG2 controls lethality in interploidy crosses of Arabidopsis. PLoS Biology *6*, 2707-2720.

Dorca-Fornell, C., Gregis, V., Grandi, V., Coupland, G., Colombo, L., and Kater, M.M. (2011). The Arabidopsis SOC1-like genes AGL42, AGL71 and AGL72 promote flowering in the shoot apical and axillary meristems. The Plant journal : for cell and molecular biology *67*, 1006-1017.

Engels, W., and Preston, C. (1979). Hybrid dysgenesis in Drosophila melanogaster: the biology of female and male sterility. Genetics *92*, 161-174.

Erilova, A., Brownfield, L., Exner, V., Rosa, M., Twell, D., Mittelsten Scheid, O., Hennig, L., and Kohler, C. (2009). Imprinting of the polycomb group gene MEDEA serves as a ploidy sensor in Arabidopsis. PLoS Genet *5*, e1000663.

Fahlgren, N., Sullivan, C.M., Kasschau, K.D., Chapman, E.J., Cumbie, J.S., Montgomery, T.A., Gilbert, S.D., Dasenko, M., Backman, T.W., Givan, S.A*., et al.* (2009). Computational and analytical framework for small RNA profiling by high-throughput sequencing. RNA *15*, 992-1002.

Feng, S., Jacobsen, S.E., and Reik, W. (2010). Epigenetic reprogramming in plant and animal development. Science *330*, 622-627.

Filson, A.J., Louvi, A., Efstratiadis, A., and Robertson, E.J. (1993). Rescue of the T-associated maternal effect in mice carrying null mutations in Igf-2 and Igf2r, two reciprocally imprinted genes. Development *118*, 731-736.

Fitz Gerald, J.N., Hui, P.S., and Berger, F. (2009). Polycomb group-dependent imprinting of the actin regulator AtFH5 regulates morphogenesis in Arabidopsis thaliana. Development *136*, 3399-3404.

Gehring, M., Bubb, K.L., and Henikoff, S. (2009). Extensive demethylation of repetitive elements during seed development underlies gene imprinting. Science *324*, 1447-1451.

Gehring, M., Huh, J.H., Hsieh, T.F., Penterman, J., Choi, Y., Harada, J.J., Goldberg, R.B., and Fischer, R.L. (2006). DEMETER DNA glycosylase establishes MEDEA polycomb gene self-imprinting by allele-specific demethylation. Cell *124*, 495-506.

Gehring, M., Missirian, V., and Henikoff, S. (2011). Genomic analysis of parent-of-origin allelic expression in Arabidopsis thaliana seeds. Plos One *6*, e23687.

Godfray, H.C., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., and Toulmin, C. (2010). Food security: the challenge of feeding 9 billion people. Science *327*, 812-818.

Goldberg, R.B., de Paiva, G., and Yadegari, R. (1994). Plant embryogenesis: zygote to seed. Science *266*, 605-614.

Grossniklaus, U., Vielle-Calzada, J.P., Hoeppner, M.A., and Gagliano, W.B. (1998). Maternal control of embryogenesis by MEDEA, a polycomb group gene in Arabidopsis. Science *280*, 446-450.

Groszmann, M., Greaves, I.K., Albertyn, Z.I., Scofield, G.N., Peacock, W.J., and Dennis, E.S. (2011). Changes in 24-nt siRNA levels in Arabidopsis hybrids suggest an epigenetic contribution to hybrid vigor. Proc Natl Acad Sci U S A *108*, 2617-2622.

Gustafson, A.M., Allen, E., Givan, S., Smith, D., Carrington, J.C., and Kasschau, K.D. (2005). ASRP: the Arabidopsis Small RNA Project Database. Nucleic Acids Res *33*, D637-640.

Gutierrez-Marcos, J.F., Pennington, P.D., Costa, L.M., and Dickinson, H.G. (2003). Imprinting in the endosperm: a possible role in preventing wide hybridization. Philos Trans R Soc Lond B Biol Sci *358*, 1105-1111.

Haag, J.R., and Pikaard, C.S. (2011). Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. Nat Rev Mol Cell Biol *12*, 483-492.

Haag, J.R., Pontes, O., and Pikaard, C.S. (2009). Metal A and metal B sites of nuclear RNA polymerases Pol IV and Pol V are required for siRNA-dependent DNA methylation and gene silencing. Plos One *4*, e4110.

Haig, D., and Westoby, M. (1989). Parent-Specific Gene Expression and the Triploid Endosperm. The American Naturalist *134*, 147-155.

Hehenberger, E., Kradolfer, D., and Kohler, C. (2012). Endosperm cellularization defines an important developmental transition for embryo development. Development *139*, 2031-2039.

Henderson, I.R., Zhang, X., Lu, C., Johnson, L., Meyers, B.C., Green, P.J., and Jacobsen, S.E. (2006). Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning. Nat Genet *38*, 721-725.

Hennig, L., and Derkacheva, M. (2009). Diversity of Polycomb group complexes in plants: same rules, different players? Trends Genet *25*, 414-423.

Herr, A.J., Jensen, M.B., Dalmay, T., and Baulcombe, D.C. (2005). RNA polymerase IV directs silencing of endogenous DNA. Science *308*, 118-120.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics *6*, 65-70.

Hsieh, T.F., Ibarra, C.A., Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R.L., and Zilberman, D. (2009). Genome-wide demethylation of Arabidopsis endosperm. Science *324*, 1451-1454.

Hsieh, T.F., Shin, J., Uzawa, R., Silva, P., Cohen, S., Bauer, M.J., Hashimoto, M., Kirkbride, R.C., Harada, J.J., Zilberman, D*., et al.* (2011). Regulation of imprinted gene expression in Arabidopsis endosperm. Proc Natl Acad Sci U S A *108*, 1755-1762.

Huettel, B., Kanno, T., Daxinger, L., Aufsatz, W., Matzke, A.J., and Matzke, M. (2006). Endogenous targets of RNA-directed DNA methylation and Pol IV in Arabidopsis. EMBO J *25*, 2828-2836.

Huh, J.H., Bauer, M.J., Hsieh, T.F., and Fischer, R. (2007). Endosperm gene imprinting and seed development. Curr Opin Genet Dev *17*, 480-485.

Ibarra, C.A., Feng, X., Schoft, V.K., Hsieh, T.F., Uzawa, R., Rodrigues, J.A., Zemach, A., Chumak, N., Machlicova, A., Nishimura, T*., et al.* (2012). Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. Science *337*, 1360-1364.

Ingouff, M., Fitz Gerald, J.N., Guerin, C., Robert, H., Sorensen, M.B., Van Damme, D., Geelen, D., Blanchoin, L., and Berger, F. (2005a). Plant formin AtFH5 is an evolutionarily conserved actin nucleator involved in cytokinesis. Nat Cell Biol *7*, 374-380.

Ingouff, M., Haseloff, J., and Berger, F. (2005b). Polycomb group genes control developmental timing of endosperm. Plant J *42*, 663-674.

Ishizu, H., Siomi, H., and Siomi, M.C. (2012). Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. Genes & Development *26*, 2361-2373.

Josefsson, C., Dilkes, B., and Comai, L. (2006). Parent-dependent loss of gene silencing during interspecies hybridization. Curr Biol *16*, 1322-1328.

Jullien, P.E., and Berger, F. (2010). Parental genome dosage imbalance deregulates imprinting in Arabidopsis. PLoS Genet *6*, e1000885.

Jullien, P.E., Katz, A., Oliva, M., Ohad, N., and Berger, F. (2006). Polycomb group complexes self-regulate imprinting of the Polycomb group gene MEDEA in Arabidopsis. Curr Biol *16*, 486-492.

Jullien, P.E., Mosquna, A., Ingouff, M., Sakata, T., Ohad, N., and Berger, F. (2008). Retinoblastoma and its binding partner MSI1 control imprinting in Arabidopsis. PLoS Biol *6*, e194.

Jullien, P.E., Susaki, D., Yelagandula, R., Higashiyama, T., and Berger, F. (2012). DNA methylation dynamics during sexual reproduction in Arabidopsis thaliana. Current Biology *22*, 1825-1830.

Kang, I.H., Steffen, J.G., Portereiko, M.F., Lloyd, A., and Drews, G.N. (2008). The AGL62 MADS domain protein regulates cellularization during endosperm development in Arabidopsis. Plant Cell *20*, 635-647.

Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., and Carrington, J.C. (2007). Genome-wide profiling and analysis of Arabidopsis siRNAs. PLoS Biol *5*, e57.

Kidwell, M. (1977). Reciprocal differences in female recombination associated with hybrid dysgenesis in Drosophila melanogaster. Genet Res *30*, 77-88.

Kidwell, M. (1981). Hybrid dysgenesis in Drosophila melanogaster: the genetics of cytotype determination in neutral strain. Genetics *98*, 275-290.

Kidwell, M. (1983). Hybrid dysgenesis in Drosophila melanogaster: factors affecting chromosomal contamination in the P-M system. Genetics *104*, 317-341.

Kinoshita, T., Miura, A., Choi, Y., Kinoshita, Y., Cao, X., Jacobsen, S.E., Fischer, R.L., and Kakutani, T. (2004). One-way control of FWA imprinting in Arabidopsis endosperm by DNA methylation. Science *303*, 521-523.

Kinoshita, Y., Saze, H., Kinoshita, T., Miura, A., Soppe, W.J.J., Koornneef, M., and Kakutani, T. (2007). Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats. Plant Journal *49*, 38-45.

Kiyosue, T., Ohad, N., Yadegari, R., Hannon, M., Dinneny, J., Wells, D., Katz, A., Margossian, L., Harada, J.J., Goldberg, R.B.*, et al.* (1999). Control of fertilization-independent endosperm development by the MEDEA polycomb gene in Arabidopsis. Proc Natl Acad Sci U S A *96*, 4186-4191.

Kohler, C., Hennig, L., Spillane, C., Pien, S., Gruissem, W., and Grossniklaus, U. (2003). The Polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene PHERES1. Genes Dev *17*, 1540-1553.

Kohler, C., Page, D.R., Gagliardini, V., and Grossniklaus, U. (2005). The Arabidopsis thaliana MEDEA Polycomb group protein controls expression of PHERES1 by parental imprinting. Nat Genet *37*, 28-30.

Kradolfer, D., Hennig, L., and Kohler, C. (2013). Increased maternal genome dosage bypasses the requirement of the FIS polycomb repressive complex 2 in Arabidopsis seed development. PLoS Genetics *9*, e1003163.

Lau, M.M., Stewart, C.E., Liu, Z., Bhatt, H., Rotwein, P., and Stewart, C.L. (1994). Loss of the imprinted IGF2/cation-independent mannose 6-phosphate receptor results in fetal overgrowth and perinatal lethality. Genes Dev *8*, 2953-2963.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet *11*, 204-220.

Le, B.H., Cheng, C., Bui, A.Q., Wagmaister, J.A., Henry, K.F., Pelletier, J., Kwong, L., Belmonte, M., Kirkbride, R., Horvath, S.*, et al.* (2010). Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proc Natl Acad Sci U S A *107*, 8063-8070.

Lee, E.K., Cibrian-Jaramillo, A., Kolokotronis, S.O., Katari, M.S., Stamatakis, A., Ott, M., Chiu, J.C., Little, D.P., Stevenson, D.W., McCombie, W.R.*, et al.* (2011). A functional phylogenomic view of the seed plants. PLoS genetics *7*, e1002411.

Lenoir, A., Lavie, L., Prieto, J.L., Goubely, C., Cote, J.C., Pelissier, T., and Deragon, J.M. (2001). The evolutionary origin and genomic organization of SINEs in *Arabidopsis thaliana*. Molecular Biology and Evolution *18*, 2315-2322.

Leroy, O., Hennig, L., Breuninger, H., Laux, T., and Kohler, C. (2007). Polycomb group proteins function in the female gametophyte to determine seed development in plants. Development *134*, 3639-3648.

Li, J., Yang, Z., Yu, B., Liu, J., and Chen, X. (2005). Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in Arabidopsis. Curr Biol *15*, 1501-1507.

Lu, C., Kulkarni, K., Souret, F.F., MuthuValliappan, R., Tej, S.S., Poethig, R.S., Henderson, I.R., Jacobsen, S.E., Wang, W., Green, P.J.*, et al.* (2006). MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. Genome Res *16*, 1276-1288.

Lu, J., Zhang, C., Baulcombe, D.C., and Chen, Z.J. (2012). Maternal siRNAs as regulators of parental genome imbalance and gene expression in endosperm of Arabidopsis seeds. Proc Natl Acad Sci U S A *109*, 5529-5534.

Luo, M., Bilodeau, P., Dennis, E.S., Peacock, W.J., and Chaudhury, A. (2000). Expression and parent-of-origin effects for FIS2, MEA, and FIE in the endosperm and embryo of developing Arabidopsis seeds. Proc Natl Acad Sci U S A *97*, 10637-10642.

Lysak, M., Fransz, P., and Schubert, I. (2006). Cytogenetic analyses of *Arabidopsis*. Methods Mol Biol *323*, 173-186.

Makarevich, G., Leroy, O., Akinci, U., Schubert, D., Clarenz, O., Goodrich, J., Grossniklaus, U., and Kohler, C. (2006). Different Polycomb group complexes regulate common target genes in Arabidopsis. EMBO Rep *7*, 947-952.

Malone, C.D., Brennecke, J., Dus, M., Stark, A., McCombie, W.R., Sachidanandam, R., and Hannon, G.J. (2009). Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. Cell *137*, 522-535.

Martienssen, R.A. (2010). Heterochromatin, small RNA and post-fertilization dysgenesis in allopolyploid and interploid hybrids of Arabidopsis. New Phytol *186*, 46-53.

Matzke, M., Kanno, T., Huettel, B., Daxinger, L., and Matzke, A.J. (2007). Targets of RNA-directed DNA methylation. Curr Opin Plant Biol *10*, 512-519.

McClintock, B. (1984). THE SIGNIFICANCE OF RESPONSES OF THE GENOME TO CHALLENGE. Science *226*, 792-801.

Meinke, D. (1994). Seed Development in *Arabidopsis thaliana*. In *Arabidopsis*, E. Meyerowitz, and C. Somerville, eds. (Cold Spring Harbor Laboratory Press), pp. 253-295.

Montgomery, T.A., Howell, M.D., Cuperus, J.T., Li, D., Hansen, J.E., Alexander, A.L., Chapman, E.J., Fahlgren, N., Allen, E., and Carrington, J.C. (2008). Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. Cell *133*, 128-141.

Moore, T., and Haig, D. (1991). Genomic imprinting in mammalian development: a parental tug-of-war. Trends Genet *7*, 45-49.

Morin, R.D., Aksay, G., Dolgosheina, E., Ebhardt, H.A., Magrini, V., Mardis, E.R., Sahinalp, S.C., and Unrau, P.J. (2008). Comparative analysis of the small RNA transcriptomes of *Pinus contorta* and *Oryza sativa*. Genome Res *18*, 571-584.

Mosher, R.A. (2011). PolIV-dependent siRNAs in plants. In Non-coding RNAs in Plants, V.A. Erdmann, and J. Barciszewski, eds. (Springer-Verlag).

Mosher, R.A., and Melnyk, C.W. (2010). siRNAs and DNA methylation: seedy epigenetics. Trends Plant Sci *15*, 204-210.

Mosher, R.A., Melnyk, C.W., Kelly, K.A., Dunn, R.M., Studholme, D.J., and Baulcombe, D.C. (2009). Uniparental expression of PolIV-dependent siRNAs in developing endosperm of Arabidopsis. Nature *460*, 283-U151.

Mosher, R.A., Schwach, F., Studholme, D., and Baulcombe, D.C. (2008). PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. Proc Natl Acad Sci U S A *105*, 3145-3150.

Mosher, R.A., Tan, E.H., Shin, J., Fischer, R.L., Pikaard, C.S., and Baulcombe, D.C. (2011). An Atypical Epigenetic Mechanism Affects Uniparental Expression of Pol IV-Dependent siRNAs. Plos One *6*, e25756.

Nam, J., Kim, J., Lee, S., An, G., Ma, H., and Nei, M. (2004). Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc Natl Acad Sci U S A *101*, 1910-1915.

Ng, D.W., Lu, J., and Chen, Z.J. (2012). Big roles for small RNAs in polyploidy, hybrid vigor, and hybrid incompatibility. Curr Opin Plant Biol *15*, 154-161.

Nobuta, K., Lu, C., Shrivastava, R., Pillay, M., De Paoli, E., Accerbi, M., Arteaga-Vazquez, M., Sidorenko, L., Jeong, D.H., Yen, Y.*, et al.* (2008). Distinct size distribution of endogeneous siRNAs in maize: Evidence from deep sequencing in the mop1-1 mutant. Proc Natl Acad Sci U S A *105*, 14958-14963.

Olsen, O.-A. (2004). Nuclear Endosperm Development in Cereals and Arabidopsis thaliana. The Plant Cell Online *16*, S214-S227.

Onodera, Y., Haag, J.R., Ream, T., Costa Nunes, P., Pontes, O., and Pikaard, C.S. (2005). Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. Cell *120*, 613-622.

Parenicova, L., de Folter, S., Kieffer, M., Horner, D.S., Favalli, C., Busscher, J., Cook, H.E., Ingram, R.M., Kater, M.M., Davies, B.*, et al.* (2003). Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: New openings to the MADS world. Plant Cell *15*, 1538-1551.

Pelisson, A. (1981). The I-R system of hybrid dysgenesis in Drosophila melanogaster: are I factor insertions responsible for the mutator effect of the I-R interaction? Mol Gen Genet *183*, 123-129.

Penterman, J., Uzawa, R., and Fischer, R.L. (2007). Genetic interactions between DNA demethylation and methylation in Arabidopsis. Plant Physiol *145*, 1549-1557.

Perry, S.E., and Wang, H. (2003). Rapid isolation of *Arabidopsis thaliana* developing embryos. BioTechniques *35*, 278-280, 282.

Pfluger, J., and Wagner, D. (2007). Histone modifications and dynamic regulation of genome accessibility in plants. Curr Opin Plant Biol *10*, 645-652.

Picard, G. (1976). Non-mendelian female sterility in Drosophila melanogaster: hereditary transmission of I factor. Genetics *83*, 107-123.

Pignatta, D., and Gehring, M. (2012). Imprinting meets genomics: new insights and new challenges. Curr Opin Plant Biol *15*, 530-535.

Pikaard, C.S., Haag, J.R., Ream, T., and Wierzbicki, A.T. (2008). Roles of RNA polymerase IV in gene silencing. Trends Plant Sci *13*, 390-397.

Pontier, D., Yahubyan, G., Vega, D., Bulski, A., Saez-Vasquez, J., Hakimi, M.A., Lerbs-Mache, S., Colot, V., and Lagrange, T. (2005). Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. Genes Dev *19*, 2030-2040.

Popp, C., Dean, W., Feng, S., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen, S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature *463*, 1101-1105.

Portereiko, M.F., Lloyd, A., Steffen, J.G., Punwani, J.A., Otsuga, D., and Drews, G.N. (2006). AGL80 is required for central cell and endosperm development in Arabidopsis. Plant Cell *18*, 1862-1872.

Raissig, M.T., Baroux, C., and Grossniklaus, U. (2011). Regulation and flexibility of genomic imprinting during seed development. Plant Cell *23*, 16-26.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nature Biotechnology *29*, 24-26.

Rubin, G., Kidwell, M., and Bingham, P. (1982). The molecular basis of P-M hybrid dysgensis: the nature of induced mutations. Cell *29*, 987-994.

Ruvkun, G. (2001). Molecular biology. Glimpses of a tiny RNA world. Science *294*, 797-799.

Schuettengruber, B., and Cavalli, G. (2009). Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. Development *136*, 3531-3542.

Scott, R.J., Spielman, M., Bailey, J., and Dickinson, H.G. (1998). Parent-of-origin effects on seed development in Arabidopsis thaliana. Development *125*, 3329-3341.

Sharma, D.R., Kaur, R., and Kumar, K. (1996). Embryo rescue in plants - a review. Euphytica *89*, 325-337.

Shirzadi, R., Andersen, E.D., Bjerkan, K.N., Gloeckle, B.M., Heese, M., Ungru, A., Winge, P., Koncz, C., Aalen, R.B., Schnittger, A.*, et al.* (2011). Genome-wide transcript profiling of endosperm without paternal contribution identifies parent-of-origin dependent regulation of AGAMOUS-LIKE36. PLoS Genet *7*, e1001303.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdzic, M., Becker, J.D., Feijo, J.A., and Martienssen, R.A. (2009). Epigenetic Reprogramming and Small RNA Silencing of Transposable Elements in Pollen. Cell *136*, 461-472.

Smith, L.M., Pontes, O., Searle, I., Yelina, N., Yousafzai, F.K., Herr, A.J., Pikaard, C.S., and Baulcombe, D.C. (2007). An SNF2 protein associated with nuclear RNA silencing and the spread of a silencing signal between cells in Arabidopsis. Plant Cell *19*, 1507-1521.

Spillane, C., MacDougall, C., Stock, C., Kohler, C., Vielle-Calzada, J.P., Nunes, S.M., Grossniklaus, U., and Goodrich, J. (2000). Interaction of the Arabidopsis polycomb group proteins FIE and MEA mediates their common phenotypes. Curr Biol *10*, 1535-1538.

Stebbins, G.L. (1976). Seeds, seedlings, and the origin of angiosperms. In Origin and Early Evolution of Angiosperms, C.B. Beck, ed. (New York, Columbia University Press), pp. 300-311.

Steffen, J.G., Kang, I.H., Portereiko, M.F., Lloyd, A., and Drews, G.N. (2008). AGL61 interacts with AGL80 and is required for central cell development in Arabidopsis. Plant Physiol *148*, 259-268.

Sunkar, R., Li, Y.F., and Jagadeeswaran, G. (2012). Functions of microRNAs in plant stress responses. Trends Plant Sci *17*, 196-203.

Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2012). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics.

Tilghman, S.M. (1999). The sins of the fathers and mothers: genomic imprinting in mammalian development. Cell *96*, 185-193.

Tiwari, S., Schulz, R., Ikeda, Y., Dytham, L., Bravo, J., Mathers, L., Spielman, M., Guzman, P., Oakey, R.J., Kinoshita, T.*, et al.* (2008). MATERNALLY EXPRESSED PAB C-TERMINAL, a novel imprinted gene in Arabidopsis, encodes the conserved C-terminal domain of polyadenylate binding proteins. Plant Cell *20*, 2387-2398.

Tiwari, S., Spielman, M., Schulz, R., Oakey, R.J., Kelsey, G., Salazar, A., Zhang, K., Pennell, R., and Scott, R.J. (2010). Transcriptional profiles underlying parent-of-origin effects in seeds of *Arabidopsis thaliana*. Bmc Plant Biology *10*, 72.

Trivers, R., and Burt, A. (1999). Kinship and genomic imprinting. Results Probl Cell Differ *25*, 1-21.

Vazquez, F. (2006). Arabidopsis endogenous small RNAs: highways and byways. Trends Plant Sci *11*, 460-468.

Villar, C.B.R., Erilova, A., Makarevich, G., Trosch, R., and Kohler, C. (2009). Control of PHERES1 Imprinting in Arabidopsis by Direct Tandem Repeats. Molecular Plant *2*, 654-660.

Walia, H., Josefsson, C., Dilkes, B., Kirkbride, R., Harada, J., and Comai, L. (2009). Dosage-dependent deregulation of an AGAMOUS-LIKE gene cluster contributes to interspecific incompatibility. Curr Biol *19*, 1128-1132.

Wang, D., Zhang, C., Hearn, D.J., Kang, I.H., Punwani, J.A., Skaggs, M.I., Drews, G.N., Schumaker, K.S., and Yadegari, R. (2010). Identification of transcription-factor genes expressed in the Arabidopsis female gametophyte. BMC plant biology *10*, 110.

Weijers, D., Geldner, N., Offringa, R., and Jurgens, G. (2001). Seed development: Early paternal gene activity in Arabidopsis. Nature *414*, 709-710.

Weinhofer, I., Hehenberger, E., Roszak, P., Hennig, L., and Kohler, C. (2010). H3K27me3 profiling of the endosperm implies exclusion of polycomb group protein targeting by DNA methylation. PLoS Genet *6*.

Wierzbicki, A.T., Haag, J.R., and Pikaard, C.S. (2008). Noncoding Transcription by RNA Polymerase Pol IVb/Pol V Mediates Transcriptional Silencing of Overlapping and Adjacent Genes. Cell *135*, 635-648.

Wierzbicki, A.T., Ream, T.S., Haag, J.R., and Pikaard, C.S. (2009). RNA polymerase V transcription guides ARGONAUTE4 to chromatin. Nat Genet *41*, 630-634.

Wolff, P., Weinhofer, I., Seguin, J., Roszak, P., Beisel, C., Donoghue, M.T., Spillane, C., Nordborg, M., Rehmsmeier, M., and Kohler, C. (2011). High-resolution analysis of parent-of-origin allelic expression in the Arabidopsis Endosperm. PLoS Genet *7*, e1002126.

Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D., Jacobsen, S.E., and Carrington, J.C. (2004). Genetic and Functional Diversification of Small RNA Pathways in Plants. PLoS Biol *2*, e104.

Yoshikawa, M., Peragine, A., Park, M.Y., and Poethig, R.S. (2005). A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. Genes & Development *19*, 2164-2175.

Yu, B., Yang, Z., Li, J., Minakhina, S., Yang, M., Padgett, R.W., Steward, R., and Chen, X. (2005). Methylation as a crucial step in plant microRNA biogenesis. Science *307*, 932-935.

Zemach, A., Kim, M.Y., Hsieh, P.H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L., and Zilberman, D. (2013). The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell *153*, 193-205.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. Science *328*, 916-919.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science *328*, 916-919.

Zhang, H., Ma, Z.-Y., Zeng, L., Tanaka, K., Zhang, C.-J., Ma, J., Bai, G., Wang, P., Zhang, S.-W., Liu, Z.-W.*, et al.* (2013). DTF1 is a core component of RNA-directed DNA methylation and may assist in the recruitment of Pol IV. Proceedings of the National Academy of Sciences *110*, 8290-8295.

Zhang, X., Henderson, I.R., Lu, C., Green, P.J., and Jacobsen, S.E. (2007). Role of RNA polymerase IV in plant small RNA metabolism. Proc Natl Acad Sci U S A *104*, 4536-4541.

Zheng, X., Pontes, O., Zhu, J., Miki, D., Zhang, F., Li, W.X., Iida, K., Kapoor, A., Pikaard, C.S., and Zhu, J.K. (2008). ROS3 is an RNA-binding protein required for DNA demethylation in Arabidopsis. Nature 2008 Oct 30;455(7217):1259-62 doi: 101038/nature07305 Epub 2008 Sep 24.

Zilberman, D., Cao, X., Johansen, L.K., Xie, Z., Carrington, J.C., and Jacobsen, S.E. (2004). Role of *Arabidopsis* ARGONAUTE4 in RNA-directed DNA methylation triggered by inverted repeats. Curr Biol *14*, 1214-1220.

Zilberman, D., Cao, X.F., and Jacobsen, S.E. (2003). ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. Science *299*, 716-719.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet *39*, 61-69.

Vita

Jie Lu was born in 1982 in Shanghai, China. In 2000, she studied at Shanghai Jiao Tong University (SJTU), China and received a Bachelor of Science degree in 2004. In 2007, she received a Master of Science degree from Fudan University, China. In August 2007, she arrived in the United States to pursue a Ph.D. degree in the Cellular and Molecular Biology program (Molecular Genetics track) at The University of Texas at Austin. In Spring 2008, she joined the Chen laboratory to study small RNAs in seed development.

Email address: lujie0823@gmail.com

This dissertation was typed by Jie Lu.